nature

www.nature.com/nature

Vol 456 | Issue no. 7223 | 11 December 2008

What next for Mars?

Future missions to the red planet require coordination — and a keen eye on costs.

The story of the Phoenix Mars lander (see page 690) raises nagging questions about how to study Earth's most interesting neighbour. Mars has been visited by spacecraft more often than any other planet. Yet the frequency of the probes has made them no cheaper. The Phoenix mission, a bargain by NASA standards, cost more than US$400 million — almost half the annual budget of the US Geological Survey that studies Earth's terrain. NASA's next planned mission to the Martian surface, a rover known as the Mars Science Laboratory (MSL), has a budget that was pushing $2 billion even before last week's announcement that its launch date was to be delayed by two years; it will now comfortably exceed that sum.

Granted, investment on this scale has produced stunning science. Mars has been mapped from orbit in exquisite detail. The remarkably long-lived Spirit and Opportunity rover missions, which landed five years ago next month, have produced a rich scientific understanding of the surface. And Phoenix itself has revealed a polar environment more dynamic than anything yet seen on the planet, with frosts coming and going, clouds dropping gentle snows and subsurface ice vanishing into vapour.

What this huge investment has not produced, however, is the long-term infrastructure and reusable technology that would make future missions more affordable. To take one example: the most difficult part of any mission to the Martian surface is landing, yet every such mission to date has used technology tailored from scratch. There are currently no plans to reuse the rocket landing system painstakingly developed for Phoenix mission, nor the air bags of the rover missions, nor even the ambitious 'sky crane' system that will supposedly lower MSL to the surface from a kind of rocket-powered hovercraft.

This constant reinvention is an indulgence that planetary exploration programmes can no longer afford. The United States and Europe, which have so far led the way in Mars exploration, should also lead the way in this respect by jointly reusing technology wherever it is remotely feasible. Rather than developing a new parachute-braking system, for example, the ExoMars rover being planned by the European Space Agency (ESA) could commit to using the air-bag technology deployed by NASA's rovers, and use as many of the other components from those missions as possible.

Looking further ahead, NASA, ESA and perhaps the new spacefaring nations in Asia should jointly start developing standard technologies for Mars exploration. One example might be a landing system designed to deliver hardy sensors to the surface at high speed, allowing the piecemeal, pay-as-you-go deployment of a planet-wide network to monitor Mars's meteorology and seismology. Another might be a shared, long-lived communications infrastructure: satellites in orbit around Mars with large antennas pointed towards Earth that are designed to relay even weak signals from the various science payloads. Such a system would mean that payloads could be built much more cheaply.

Meanwhile, the most effective way to improve the returns on Mars exploration would be better cost discipline. Missions understandably go over budget. There is always pressure to include more instruments, not least because researchers are aware of how few opportunities there are. And the incentives to be realistic in initial budgeting are relatively small, as the missions are unlikely to be cancelled if the costs escalate.

Cancelling, or radically downscaling, overbudget missions such as MSL would set science back in the short term. It would also be difficult politically, as aerospace contractors and NASA's spacecraft-building centres have a lot of clout. But the only way to make Mars exploration a more regular affair is to stop the missions from costing too much. Eventually, a virtuous circle could be established: cheaper missions would mean more of them, which would mean less pressure to overload each one, in turn keeping the costs down. The fact that a series of failures followed an attempt to do this in the 1990s does not mean it was the wrong thing to do; it means that the effort needs clearer management and accountability, technology that can be reused from mission to mission and patience.

The hard truth is that trying to understand one planet from the surface of another is going to be a long, drawn-out affair. The early history of the space age, in which ancient points of light turned overnight into worlds as complex as our own — accompanied by detailed and often spectacular images — promised an unsustainable pace of progress. The progress of planetary science now demands time and a sound strategy. ■

Watching Big Brother

The world is sleepwalking into a surveillance society. A European court ruling offers a timely wake-up call.

The European Court of Human Rights last week issued an opinion to which the developers and users of new technologies should pay heed. "Any state claiming a pioneer role in the development of new technologies," the court said, bears special responsibility for "carefully balancing the potential benefits of the extensive use of such techniques against important private-life interests."

This timely reminder, coming just before the 60th anniversary on 10 December of the Universal Declaration of Human Rights, was a strike at the British government's policy of keeping DNA fingerprint records on an ever-growing number of innocent people in England and Wales. That policy was indeed pioneering: the United Kingdom authorized the world's first national DNA database in 1994. The database has since proved helpful in solving a small but significant number of crimes, and many countries have followed suit. But the system, originally intended to cover only people convicted of serious

offences such as murder and rape, has expanded to include samples from anyone arrested for any recordable offence — even dropping litter — and to keep them on record even if the arrest does not lead to a conviction. Moreover, the data have been used not just to match individuals to crimes, as originally intended, but also for more dubious applications such as searching for a perpetrator's blood relatives (see *Nature* **449,** 377–378; 2007).

As a result, Britain's DNA database has grown to become proportionally the largest in the world; it contains samples and data for 7% of UK citizens, far ahead of Austria in second place with 1%. The European Court's decision thus paves the way for 850,000 people, including tens of thousands of children, to have their records removed.

DNA databases are but one small tip of the emerging surveillance society. Even leaving aside law-enforcement and security initiatives, vast amounts of data are being collected by private firms through citizens' use of credit cards, mobile telephones and electronic travel tickets, not to mention the Internet and e-mail. These data are typically gathered not for any sinister purpose, but as legitimate efforts to offer customers better service. But the databases exist. And without strong safeguards, they could slowly and steadily be linked into an all-pervasive monitoring system that would make George Orwell's concept of 1984 look technologically tame — all in the name of security, efficiency and convenience.

Such concerns are certainly not new; Orwell's book was published in 1949. But the dizzying pace of technological advance makes them ever more salient — even as it makes the world's multitude of existing privacy acts seem light-years behind. Scientists, in particular, have an ongoing responsibility to reflect on the human-rights issues raised by the technologies they develop, and to lobby for appropriate oversight and controls. The risks posed by overzealous surveillance (see page 680) and the associated technologies are topics that should be addressed by the American Association for the Advancement of Science's Science and Human Rights Coalition, a forum of scientific bodies and human-rights groups to be created in January 2009 (see *Nature* **456,** 2; 2008).

"Without strong safeguards, databases could slowly and steadily be linked into an all-pervasive monitoring system."

Technology can be a powerful force for human rights. Earth-observation satellites, for example, have provided evidence on conflicts and ethnic atrocities in areas where journalists are banned. And DNA fingerprinting has resulted in the freeing of wrongly convicted individuals, a role exemplified by the US Innocence Project in New York. The idea that the identity of a human can be revealed from samples of any cell in his or her body is a symbol of the fact that every person is unique. The declaration of human rights asks us to treasure and honour all these unique individuals with respect for their autonomy — not to simply look for better ways to barcode them. ■

Failure in the field

The US military's human-terrain programme needs to be brought to a swift close.

The US Department of Defense's Human Terrain System, an attempt to have social sciences inform military decision-making, is failing on every level.

In theory, it is a good idea. The Human Terrain System aims to embed anthropologists and other social scientists in military units in Iraq and Afghanistan to help improve understanding of local cultures and thus relieve tensions between civilians and soldiers. In practice, however, it has been a disaster. Questions have been raised about how well the programme vets its employees (see *Nature* **455,** 583–585; 2008). Some scientists who have joined the system have complained about inadequate training. And qualified researchers have been dismissed for seemingly trivial reasons, even though much more questionable people seem to breeze onto the payroll.

A case in point is Issan Hamama. Under investigation by the Federal Bureau of Investigation since 2003 as a possible former spy for Saddam Hussein, Hamama nonetheless managed to secure a job as a translator for the Human Terrain System. Late last month, he was arrested in Maine and indicted for conspiracy; he is currently free on bail.

Another contractor, bodyguard Don Ayala, is also out on bail after being indicted for a murder committed in Afghanistan last month. According to a military affidavit, Ayala shot Abdul Salam at close range in the head after Salam doused his colleague, social scientist Paula Lloyd, with petrol and set her on fire. Lloyd had approached Salam on the street — he was carrying a fuel jar — to ask him about the price of petrol.

Lloyd returned to the United States to recover from her burns; some of her colleagues have not been so lucky. Social scientist Michael Bhatia was killed in Afghanistan in May; Nicole Suveges, a PhD student from Johns Hopkins University in Baltimore, Maryland, died in Iraq the following month.

Their names and sacrifices should be remembered. But the programme that employed them should not — except, perhaps, as an example of yet another good idea gone wrong on the war fields of Iraq and Afghanistan.

The immediate problems with the Human Terrain System can be traced to BAE Systems, the military contractor based in Rockville, Maryland, that screens potential employees, then trains those it hires. It has failed in every one of those functions, and army management has failed in its oversight of BAE.

But the larger question is whether the Human Terrain System is viable at all. *Nature* is not opposed in principle to academics working with the military; we have said before that social science can and should inform military policy (see *Nature* **454,** 138; 2008). We continue to believe that the insights of science have much to offer strategies in a war zone — not least through training combat troops to understand the local cultures within which they operate.

But as currently constituted, the Human Terrain System is not the way to do this. Unless the programme can be reborn in a format less plagued by deadly mistakes, it needs to be closed down. ■

RESEARCH HIGHLIGHTS

ASTRONOMY

That's the weigh to do it

Astrophys. J. **689,** 1044–1062 (2008)

The star S0-2 orbits very close to the black hole at the centre of the Milky Way. More than a decade's worth of data from the Keck telescope in Hawaii has allowed Andrea Ghez of the University of California, Los Angeles, and her colleagues to produce a model of this star's orbit around the black hole that is by far the most accurate to date.

Using measurements of S0-2's movement back and forth along a straight line between it and Earth, and also of its movements from side to side, the researchers estimate that the black hole, if stationary, has a mass of between 4 million and 5 million times that of the Sun, and is 8–8.9 kiloparsecs from Earth.

DEVELOPMENTAL BIOLOGY

Swing time

Development **136,** 129–138 (2009)

All mammalian embryos grow up female unless told not to, with their 'supporting' cells turning into granulosa cells, which occur in the ovaries. Only when a Y-chromosome gene called *Sry* triggers the supporters to specialize as testicular cells that nurture growing sperm — Sertoli cells — does maleness follow.

A team led by Yoshiakira Kanai of the University of Tokyo has created a line of transgenic mice in which the turning on and off of *Sry* can be very tightly controlled. Using these mice, they have shown that *Sry* activation must occur during a six-hour window eleven days after mating if the adult is to have Sertoli rather than granulosa cells.

M. DAVIS

Climate fallout

Geophys. Res. Lett. doi:10.1029/2008GL035556 (2008)

Most glaciers have a radioactive layer corresponding to the period of atmospheric nuclear testing in the 1950s and 1960s. The Naimona'nyi glacier, which sits six kilometres above sea level on the Tibetan mountain of the same name (pictured), lacks this fallout layer, according to ice cores drilled by Natalie Kehrwald of Ohio State University in Columbus and her colleagues. This suggests that the glacier has accumulated no new ice for more than 50 years.

It is therefore the highest glacier reported to be annually losing mass. If the same is true for other, similarly situated glaciers in the region, the downstream consequences for people who depend on the glaciers' meltwater — a significant fraction of Asia's population — could be severe.

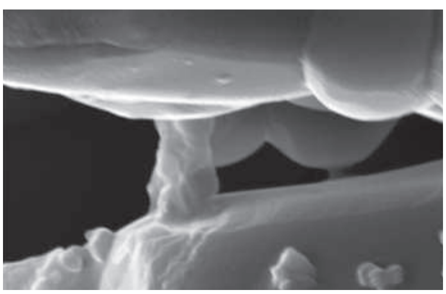
MATERIALS SCIENCE

Tough shell secrets

Science **322,** 1516–1520 (2008)

Nacre, the substance that forms the shell of many molluscs, is an extremely strong material, even though it is made only of brittle and soft components. This has inspired researchers at the Lawrence Berkeley National Laboratory in California to create a material with a similar architecture to nacre's but such superior toughness that its strength is comparable to that of aluminium alloys.

Robert Ritchie and his co-workers noted that the secret to nacre's useful properties lies in the stacking of its layers, which stops cracks from propagating. They recreated this feature using aluminium oxide and an organic polymer. The polymer was grafted (pictured below left) onto micrometre- and nanometre-scale ceramic pieces to help the two substances stick together.



E. MUNCH *ET AL.*

MOLECULAR BIOLOGY

Not so common sense

Science doi:10.1126/science.1162253; doi: 10.1126/science.1162228 (2008)

A surprising number of the proteins that transcribe DNA into RNA are reading the DNA backwards, two large-scale surveys report.

RNA polymerases use DNA templates to make RNA molecules, some of which are then used to make proteins. Phillip Sharp of the Massachusetts Institute of Technology in Cambridge and his colleagues have found that, whereas one type of polymerase binds to the start of a gene and proceeds towards the end, another often binds near the start site and heads in the opposite direction. This creates a shorter, 'antisense' (complementary) RNA molecule.

Meanwhile, John Lis and his colleagues at Cornell University in Ithaca, New York, reached a similar conclusion after creating a quantitative map of sites at which polymerase is bound to DNA and is actively synthesizing RNA. The function of these short antisense RNAs remains unclear.

MICROBIOLOGY

The ABC of anthrax

PLoS Pathog. **4,** e1000210 (2008)

The bacterium that causes anthrax, *Bacillus anthracis*, is a quiet little spore — until it gets into a mammal. Then it produces a host-killing toxin and self-protecting capsules in response to higher bicarbonate levels than

normally occur in the environment. How it recognizes this compound was the subject of a recent study by Marta Perego and her colleagues at the Scripps Research Institute in La Jolla, California.

They searched the literature for other organisms that use bicarbonate, and found work on a photosynthetic bacterium, *Synechococcus elongatus*. This shuttles bicarbonate through its cell membrane by means of a multiprotein structure known as an ABC transporter. A scan of *B. anthracis*' genome for genes that encode similar proteins pinpointed a section named BAS2714-12. Deleting this region rendered the pathogen harmless to mice.

ECOLOGY

The alien within

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0809423105 (2008)

South American fire ants (*Solenopsis invicta*) have been shown to invade Florida's extensive pine flatwoods only where human activities such as mowing or ploughing create a disturbance.

Joshua King and Walter Tschinkel of Florida State University in Tallahassee set up several types of plot: undisturbed, mowed and ploughed, and added fire ants to some of each plot type. Fire ants can't displace native ants on their own; instead, disturbance makes this possible by reducing native ant populations and creating habitats in which fire ants do particularly well.

NANOTECHNOLOGY

Scalable solution

Nature Nanotech. doi:10.1038/nnano.2008.365 (2008)

Sodium; ethanol; heat; sound waves: those easy-to-come-by, cheap ingredients are all that John Stride and his team at the University of New South Wales in Australia needed to make a couple of grams of pure carbon in the form of graphene.

Graphene is just one atom thick and looks like molecular chicken wire. First isolated in 2004, it is the strongest material known, and should make tomorrow's mix of power-hungry consumer gadgets lighter and more efficient. The problem has been that graphene is devilishly difficult to manufacture in large amounts.

Stride and his colleagues first reacted ethanol and sodium, then broke down a solid intermediate of the reaction by heating. This generated a fused pack of graphene sheets that they then broke apart with gentle sonication.

ANIMAL BEHAVIOUR

Love bite

J. Evol. Biol. doi:10.1111/j.1420-9101.2008.01636.x (2008)

Female sagebrush crickets (*Cyphoderris strepitans*) munch off males' wingtips during mating (pictured), stimulating an immune reaction that saps males' energy and makes it harder for them to woo other females.

Scott Sakaluk and his students at Illinois State University in the town of Normal arrived at this conclusion after measuring the cost of the immune response to male sexual vigour. They injected some wild-caught virgin males with lipopolysaccharides that trigger their immune systems, and others with a substance that has no such effect.

Lipopolysaccharides drastically reduced the courtship behaviour and mate procurement of male sagebrush crickets. Females seem to gain nothing more than a good meal from this act of aggression.

D. FUNK

IMMUNOLOGY

T cells get tired

Nature Immunol. doi:10.1038/ni.1679 (2008)

During chronic infection, immune systems' T cells often lose their ability to fight pathogens. Research now shows that this T-cell exhaustion is caused by complex negative-feedback pathways that depend on the types of inhibitory receptor the cells express on their surfaces.

By examining gene expression in mice with a chronic infection, John Wherry of the Wistar Institute in Philadelphia, Pennsylvania, and his colleagues found that up to seven inhibitory receptors, which are not made in fully functional T cells, are expressed in exhausted T cells. The more severe the infection, the more members and amounts of this septet the T cells produced.

Simultaneously blocking the function of two such receptors — PD-1 and LAG-3 — incompletely restored exhausted T cells' activity, indicating that multiple pathways are involved in T-cell exhaustion.

JOURNAL CLUB

Eric D. Tytell
University of Maryland, College Park

A neuroscientist marvels at our ability to learn unnatural tasks.

I find driving mind-boggling. As a neuroscientist studying motor control, I am amazed that nervous systems can adapt to the unnatural demands of operating a car. After all, humans did not evolve in habitats with steering wheels or accelerator pedals. What makes our ability to drive so curious is that it requires the modification of reflexes — twisting the steering wheel, for instance, rather than jumping aside, when an obstacle approaches.

Mark Wagner and Maurice Smith have shed some light on this curiosity. They show that the brain generalizes unnatural physical regimes, such as driving, to produce an appropriate corrective response to an unexpected change, even when that change has not been met before (M. J. Wagner and M. A. Smith *J. Neurosci.* **28,** 10663–10673; 2008).

The duo trained undergraduates to reach quickly for a target with one hand while holding on to a motorized arm with the other. The faster the students reached, the stronger the motorized arm pushed them off course.

Initially, the students made large errors, but they soon compensated for the lateral forces. Were their brains learning the dynamics of the new force, though, or were they reassigning the activation of muscles in the spinal cord from those for reaching towards those that normally help to generate sideways pushes?

Surprising the students with a sudden pulse of force in the reaching direction provided an answer. They compensated with almost ideal corrective forces, which spinal reflexes alone could not have achieved. The slight delay in the students' responses also indicates that their brains were working from an internal model of the new force regime. How the brain develops such a model is unknown, but this paper should drive that research.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Malaria vaccine gets shot in the arm from tests

Promising results pave the way for a vaccine candidate to undergo full-blown trials across Africa.

A long-anticipated candidate vaccine for malaria may be more than 50% effective at preventing the disease in African children, say results published this week in *The New England Journal of Medicine*[1,2]. The work sets the scene for phase III clinical trials to start within the next few months.

The vaccine, called RTS,S, has been in development for more than two decades[3], and earlier studies had suggested that it was effective at rates of little better than 30%. Yet even modest protection against malaria is deemed a success because of the disease's deadly and endemic nature. But "there's something quite important in getting beyond 50%", says Kevin Marsh, director of the KEMRI–Wellcome research programme in Kilifi, Kenya. "It really strengthens the argument that it will be possible to develop something like a complete vaccine."

Malaria experts are impressed with the latest findings, which were also presented on 8 December at a meeting of the American Society of Tropical Medicine and Hygiene in New Orleans, Louisiana. "Overall, these show good progress with RTS,S," says Adrian Hill, director of the Jenner Institute in Oxford, UK, who was not directly involved in the work.

The results are from phase II trials conducted at three sites in western Africa by the vaccine developer GlaxoSmithKline (GSK) and a wide-ranging team of clinical, academic and non-profit funding partners. The next step will be phase III trials, which scale up the numbers treated. These will be funded largely by the Bill & Melinda Gates Foundation in Seattle, and will involve up to 16,000 children at 10 sites in 7 African countries (see map).

In work described in one of the new papers[1], more than 800 children aged 5 to 17 months in Kenya and Tanzania received either the RTS,S vaccine or a rabies vaccine as a control. After an average follow-up of eight months, only 32 children in the RTS,S group had developed clinical malaria compared with 66 in the control group. That corresponds to roughly 53% efficacy.

The trial tested a newer vaccine formulation that uses an adjuvant called AS01 to help stimulate an immune response. AS01 seems to trigger much stronger antibody responses in the body than the previous adjuvant used, but this still doesn't translate into massive gains in protection against clinical malaria. Marsh, a co-author of the work, calls this disparity puzzling and frustrating, as complicated clinical trials such as these are the only way to get a good idea of efficacy. Part of the problem, Marsh says, is that "no one really knows how RTS,S works".

The other study, conducted in Bagamoyo, Tanzania, was smaller and was mainly concerned with the feasibility and safety of administering RTS,S with other vaccines. A group of 340 infants was given either RTS,S or the hepatitis B vaccine, together with routine vaccines for diphtheria, tetanus, pertussis and

France cracks down on Iranian scientists

Europe's largest basic-research agency, the CNRS in France, is now vetting not only non-European Union (EU) researchers working on sensitive topics, but also all scientists from Iran regardless of what they work on. Researchers allege that the move exacerbates an already difficult situation in which visa applications are denied based on nationality.

A 12 November memo, relayed by the CNRS's security office from the state national security and industrial espionage services, requires that all projects with Iran go through formal security screening, to identify ones potentially meriting further scrutiny. It translates UN sanctions that call for member states not to provide Iran with technical assistance or training that might be used for the proliferation of weapons of mass destruction, says Joseph Illand, who heads the CNRS security office.

It's the first case Illand can recall of such an explicit country-specific instruction at the CNRS — although similar, if unwritten, restrictions have in practice long been in effect for Iran and other countries considered as security risks. Given the threat of proliferation, such extra scrutiny is normal, he says.

R. HOMAVANDI/REUTERS

The rules expand on those introduced in January 2005, which require extra scrutiny for visits from non-EU scientists whose work relates to an undisclosed list of 'sensitive' laboratories or topics. In such cases, the heads of host laboratories must give details of proposed visits, and their opinion of the visitor, to security authorities. For less-sensitive laboratories, lab heads need only declare foreign visitors monthly, whereas no-risk labs must simply maintain an internal register that can be consulted by authorities if needed.

In recent years, Iranian researchers have struggled for academic equality and scientific access. In July, for instance, the Netherlands banned Iranians from accessing nuclear laboratories or courses, and the United States has subjected Iranian visa requests to lengthy interagency reviews in the wake of the terrorist attacks of 11 September 2001. Britain last year also introduced tough vetting of non-EU science students.

The fact that France is cracking down on Iranian researchers

J.-M. MAAS/DARBY COMMUNICATIONS

Two African studies raise hopes that an effective malaria vaccine may be in sight.

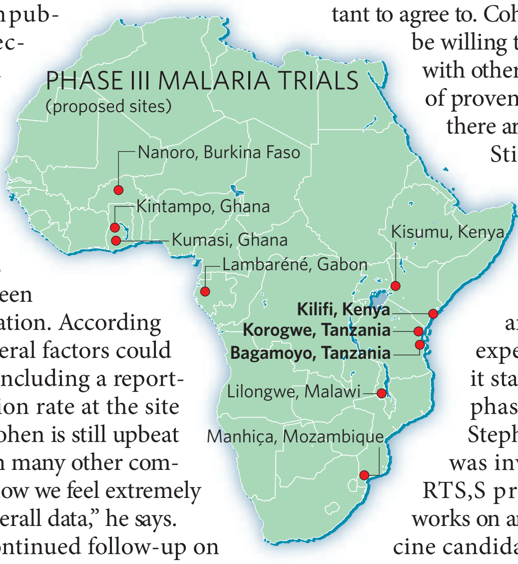
Haemophilus influenza type b, offered under a World Health Organization immunization programme. "Using the vehicle of distribution that already exists would make it easy to rapidly scale up the accessibility of a potentially good malaria vaccine to many children," says first author Salim Abdulla of the Ifakara Health Institute in Bagamoyo.

RTS,S, used in this study with the older adjuvant AS02, did not interfere with the other vaccines; nor did these other vaccines interfere with the ability of RTS,S to elicit an immune response. Although the study was not designed to test for effectiveness, the results did suggest that malaria infection rates were lower in vaccinated children.

Questions remain about how long protection might last. A study[4] in Mozambique published in 2004 found that the RTS,S vaccine was more than 30% effective in children aged 1 to 4 years. Long-term follow-up in one cohort showed that that efficacy was maintained, but unpublished data on a second cohort showed a dramatic fall in the 18 months following vaccination.

Joe Cohen, a vice-president at GSK and a key developer of RTS,S, says these findings have been submitted for publication. According to other sources, several factors could explain the fall-off, including a reportedly high transmission rate at the site of cohort two. But Cohen is still upbeat about RTS,S based on many other completed trials. "Right now we feel extremely confident with the overall data," he says.

New trials, and continued follow-up on current studies, should help pin down exactly for how long efficacy might be maintained. The phase III trials will be conducted in areas of varying transmission rates, which may affect the observed efficacy. And in general, sites with ongoing malaria trials tend to have lower transmission rates than other parts of Africa, simply because countermeasures such as bednets are more widely available at these developed research sites.



Some researchers, including Hill, want greater attention paid to developing a more effective second-generation vaccine. This might combine RTS,S with other promising candidates, which he says GSK has been reluctant to agree to. Cohen says GSK would be willing to try combinations with other vaccine candidates of proven efficacy, but so far there aren't any.

Still, the new results have edged RTS,S closer to its goal. "The vaccine is clearly having an effect, and we're all waiting expectantly to see how it stands up in the large phase III studies," says Stephen Hoffman, who was involved developing RTS,S precursors but now works on another malaria vaccine candidate at his company Sanaria in Rockville, Maryland. "It's an exciting time."

Brendan Maher

1. Bejon, P. *et al. N. Engl. J. Med.* **359,** 2521–2532 (2008).
2. Abdulla, S. *et al. N. Engl. J. Med.* **359,** 2533–2544 (2008).
3. Maher, B. *Nature* **451,** 1042–1046 (2008).
4. Alonso, P. L. *et al. Lancet* **364,** 1411–1420 (2004).

---

doesn't surprise Carthage Smith, deputy executive director for the International Council for Science in Paris, and a member of its human-rights committee. "There are legitimate concerns in the international community at present about the risk of technology transfer to Iran and other specific countries," he says. "We can't ignore that."

But he worries about how the rules might be implemented. "Any vetting must be separate from country lists, with each individual case being judged on its merits," he says. Otherwise, overzealous bureaucrats could turn country listings into effective blacklists. Formalizing extra vetting for Iranian researchers may have secondary effects, admits Illand, in that French hosts might show less interest in collaborating with Iranians to avoid administrative hassle, or government bureaucrats may be tempted to play safe and turn down visa applications. Many within the CNRS and the French administration, he says, support greater collaboration with Iran as a form of scientific diplomacy.

Overall statistics are not available, but *Nature* has gathered complaints from more than a dozen Iranian scientists who allege that French authorities recently either refused their requests for visas or simply never responded.

"They need to stop looking at researchers' passports, and start looking at their work."

Take Ali Tahzibi, a mathematician and Iranian national at the Federal University of São Carlos in Brazil. In July he applied for a visa to work for a year, starting in October, with Christian Bonatti at the CNRS's Institute of Mathematics, part of the University of Bourgogne in Dijon. Tahzibi took a year's sabbatical, bought plane tickets for himself and his family, rented an apartment in Dijon, and arranged schooling for his son. The visa never materialized. "To wait up until just hours before your flight in the hope of getting a visa without obtaining any response was torture for a family with a little child," says Tahzibi.

Visa-seeking researchers share common stories of being passed back and forth within the French bureaucracy. "It's a typical administrative vicious circle," says Pierre Sens, a researcher at the Paris Industrial Physics and Chemistry Higher Educational Institution, who tried to intervene on behalf of a former postdoc and Iranian national. The office of Valérie Pécresse, the minister of research and higher education, declined to comment on the allegations or their impact on collaboration, instead referring *Nature* to the ministry of immigration.

The cases and the CNRS policy are "absolutely astonishing, if not scandalous on the part of a country such as France, which sees itself as the cradle of human rights", says Thierry Dauxois, a CNRS physicist at the Ecole Normale Supérieure in Lyon. "They need to stop looking at researchers' passports, and start looking at their work."

Declan Butler

See also Editorial, page 675.

GOT A NEWS TIP?
Send any article ideas for Nature's News section to **newstips@nature.com**

K. CAMPBELL/GETTY

the Freed paper "so important that it went through intense vetting".

Clashes between academics and conservation managers are not uncommon — but rarely do relations become quite so strained. Since 2006, Freed has not been granted permits to work in the reserve because of the ongoing disputes. When Freed asked graduate students to follow the birds instead, Richard Waas, the refuge manager at the time, demanded that they stop, saying they did not have the necessary permits. Refuge officials then threatened to call law-enforcement rangers.

Blocked projects

Dawn Reding, now a PhD candidate at Iowa State University in Ames, is one of several students who have been affected by the situation. She says that being refused permission for her akepa study in 2004 delayed her master's project by a year; "I was really upset," she says.

Matthew Medeiros, a native Hawaiian and undergraduate student, says that his proposal to use artificial nests to study akepa was rejected because of the conflict. "If you were associated with [Freed], Jeffrey blocked everything," says Medeiros, who is now a graduate student at the University of Missouri in St Louis.

Jeffrey says the student cases have been "blown out of proportion".

For years Freed and refuge officials also argued over how best to assess blood samples from akepa to test them for malaria. The fight dragged on for so long that key work was approved only after Freed's federal grant for the work had expired.

Gary Ostrander, the vice-chancellor for research at the University of Hawaii, says he hopes the university can "continue to work collectively" with the refuge. ■

Rex Dalton

opposed the proposal, arguing that it was prejudicial to basic scientists. "It's a terrible idea," said Thomas Kelly, director of the Sloan-Kettering Institute in New York City. "It invites gaming the system. You check that [human subjects] box and get 50% more space."

Mary-Claire King, a geneticist at the University of Washington in Seattle, argued that applications involving human subjects do not require any more space than those for "flies, worms, mice, bacteria or yeast".

The decision of whether or not to implement the change ultimately rests with Kington, who says it will be made "pretty quickly". Whatever the outcome, the new rules won't come into effect until 2010. ■

Meredith Wadman

Plant hormone study pulled

Two years ago, three papers[1–3] made a splash in the plant biology world by identifying protein receptors for abscisic acid (ABA), a key hormone in plant physiology. This week, *Nature* is retracting one of those papers[1] after reports that the work could not be replicated. The two other papers still stand, although one[3] has been questioned by several researchers[4–7].

The retraction is a setback in the search to find receptors for ABA, a notoriously difficult task. ABA responds to environmental stresses such as drought, and is an alluring target for agricultural companies hoping to produce drought-resistant crops.

The retraction in *Nature* comes from the team that did the work, led by Robert Hill of the University of Manitoba in Winnipeg, Canada. Hill says the experiments were carried out by postdoc Fawzi Razem, who has since left the university; Hill says he does not know for where. A *Nature* reporter was unable to contact Razem.

John Danakas, director of public affairs for the university, says that he cannot comment on this case in particular, but says that "normally when there are problems with research results being reproducible, there would be an investigation at the senior administrative level".

According to the search tool Scopus, Hill's paper has been cited 120 times, and is the most highly cited study among the 95 results for 'abscisic acid receptor' in the past three years.

J. BURGESS/SPL

Protein isolation can be tricky in plants, such as thale cress.

Hill says he moved on from the project after the paper was published, but started to realize something might be wrong after some of his postdocs returned to it in late 2007 and could not replicate the results. He was also contacted by Catherine Day of the University of Otago in Dunedin, New Zealand, who could not repeat the work either. "I suspect it has hurt a lot of people," he says. "It certainly hurt my lab quite significantly. I certainly feel very bad about it and am very apologetic to the community."

According to Hill, his team interpreted an experiment as showing that ABA broke up a binding interaction between the receptor candidate FCA, involved in flowering, and another protein, FY, because ABA bound better to FCA than did FY. In fact, he says, the FY protein was not properly prepared, and probably never bound to FCA at all. Another experiment purporting to show direct binding between FCA and ABA was miscalculated and, says Hill, "the assay procedure is also suspect".

Some plant-hormone specialists had wondered about the results of this and the other papers from the beginning, as they did not fit with the rest of what was known about ABA. The studies all used *in vitro* biochemical techniques to identify the receptor, rather than the more usual method of identifying and cloning mutated genes before determining if they coded for a receptor.

"There was nothing really wrong with the paper; you couldn't say 'this gel isn't any good'," says Peter McCourt, an ABA specialist at the University of Toronto, Canada. But "none of it linked up with any of the genetics that had been done in the previous 20 years".

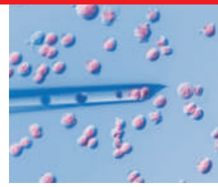
The other papers identify similarly unexpected receptors. "People who are doing this are sceptical and are not putting all their money on these findings," says Jianhua Zhang, who works on other aspects of ABA at Hong Kong Baptist University.

Day blames the problems on tricky assays. "I think they've just had a bad start. These proteins are difficult to work with." McCourt thinks such experiments might be too complicated to pursue. Collecting proteins from plant cells for binding studies is easier said than done, he says, as plant cells are particularly crowded, complex places.

Meanwhile, Hill says, "I just want to get this behind me." ■

Emma Marris

1. Razem, F. A. *et al.* *Nature* **439,** 290–294 (2006).
2. Shen, Y. Y. *et al.* *Nature* **443,** 823–826 (2006).
3. Liu, X. *et al.* *Science* **315,** 1712–1716 (2007).
4. Johnston, C. *et al.* *Science* **318,** 914 (2007).
5. Guo, J. *et al.* *PLoS ONE* **3** (8): e2982 (2008).
6. Gao, Y. *et al.* *Plant J.* **52,** 1001–1013 (2007).
7. Illingworth, C. J. *et al.* *Biophys. Chem.* **133,** 28–35 (2008).



ALAMY

DUBIOUS DISPENSARIES
Stem cell society urges government to close bogus clinics
www.nature.com/news

Patent pledge to Indian universities

Critics worry that the push for technology transfer is moving too fast.

India's parliament will soon be scrutinizing a bill intended to help publicly funded institutes and universities commercialize their research. India's science minister, Kapil Sibal, is confident that the bill will become law and will help Indian universities "make millions through patents".

But the bill has been pushed through without an open debate and is not publicly available, prompting concern. "In a democracy you cannot bulldoze a bill like this," says Pushpa Bhargava, former vice-chair of the National Knowledge Commission.

M. QURAISHI/AP

Kapil Sibal, India's science minister, wants the universities to create wealth.

The idea is widely believed to be the brainchild of Raghunath Mashelkar, former director-general of the Council of Scientific and Industrial Research (CSIR), who has turned his own labs into veritable patent factories (see *Nature* **442,** 120; 2006). After the knowledge commission backed the move in 2007, the government asked the Department of Biotechnology to draft wording for the bill, which the Indian cabinet approved on 31 October. It had been slated to go before a parliamentary committee this month, but the Mumbai terror attacks may cause it to be pushed into early 2009.

The bill is modelled on the 1980 US Bayh–Dole Act, which allowed US universities to patent discoveries derived from federally funded work. According to the Indian bill, scientists would be allowed to retain 30% of the net income earned from patents and licences. The scientist's institute would retain 40%, with the rest going into a fund maintained by the institute for managing intellectual property. Researchers in publicly funded institutes or universities would also be allowed, for the first time, to set up and work in private companies without having to leave their academic jobs.

"The benefits of publicly funded research are not reaching the public," Sibal told a meeting of the Federation of Indian Chambers of Commerce and Industry (FICCI) in October. "We need to create an environment in which wealth can be generated from the university system."

Only a few top institutes in India have a good grasp of patents and markets. Between 1990 and 2002, Indian universities accounted for only 3% of patents filed by Indian organizations in the Indian patent office, and 1% of Indian patents filed in the US patent office, according to the New Delhi-based National Institute of Science, Technology and Development Studies. Researchers at institutes without a well-developed technology-transfer office can apply to the National Research Development Corporation in New Delhi for help in commercializing their work.

Heads of publicly funded institutes support the bill in principle. "This will empower universities in an era of patents and competitiveness," says Samir Brahmachari, director general of the CSIR. But he notes that scientists and institutes also need to strike a balance between earning profits and keeping the public interest in mind, especially in fields such as affordable health care and infectious diseases.

Most university academics have not read ▶

India creates funding council for basic science

The Indian government plans to create an independent agency to promote basic research in science and engineering, along the lines of the US National Science Foundation, prime minister Manmohan Singh announced last week.

The new body, to be called the National Science and Engineering Research Board (NSERB), is expected to control an annual budget of 10 billion rupees (US$200 million) annually — equivalent to about 15% of last year's total funding for science — in new funding. "It will provide unfettered financial assistance to researchers, academic institutions, research laboratories and industrial concerns," Singh told scientists on 3 December.

The idea of an autonomous funding agency was first mooted by the Department of Science and Technology (DST) in 2002. Prominent scientists including C. N. R. Rao, chairman of the prime minister's scientific advisory council, have been lobbying for it since. "I am happy that it has finally come through," says Rao.

Parliament will establish the NSERB during its session beginning this week, and the agency will be functional from the financial year beginning 1 April 2009, says DST secretary Thirumalachari Ramasamy. Its nine-member board will be chaired by an eminent scientist, to be named by the cabinet, and will have several advisory panels.

Scientists can currently get funding from departments such as health, agriculture, energy or biotechnology — but usually only for research related to these areas. The biggest source of funds for basic science is the DST's Science & Engineering Research Council (SERC), which was set up in 1974. Last year, it distributed 3.6 billion rupees for projects in all disciplines.

Rao says that the current funding mechanisms are too bureaucratic. Ramasamy says that "the NSERB can be fast" as it is not subject to the same government rules as SERC. SERC would continue to function but only to provide money to improve university science laboratories and to handle scholarship and fellowship programmes.

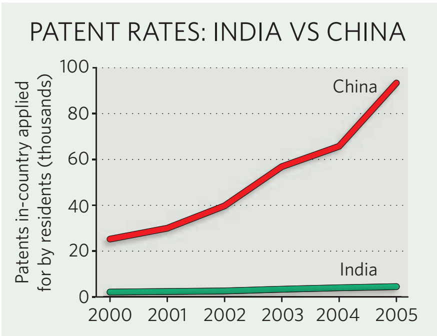
The idea of an autonomous board is "wonderful", says Seyed Hasnain, vice-chancellor of the University of Hyderabad and a member of the prime minister's scientific advisory council. "My only concern is it should not end up as another agency funding scientists in institutions already having a big budget. The bulk of the board's funding must go to university researchers, who are most neglected now." ■

K. S. Jayaraman

the bill, and their opinions on the initiative are mixed. Some welcome it, saying it will increase awareness of patenting. Others are concerned that teaching and basic research might get sidelined.

Indian industry is enthusiastic about the move. Amit Mitra, the FICCI's secretary general, argues that the Bayh–Dole Act helped US universities earn revenue from key discoveries, such as the anti-AIDS drug stavudine at Yale University and recombinant-DNA technology at the University of California and Stanford University.

But some analysts say that the US model cannot be replicated in India, where most universities are hampered by poor funds and outdated infrastructure, and where fewer students are choosing to take up science. "Many students end up doing neither frontline basic science nor applied technology, and fall in-between with mediocre work," notes Ashok Parthasarathi, former science adviser to the Indian government. He says the bill is both "impractical and



SOURCE: WIPO

infeasible" to implement, specifying what he calls "absurd time periods". The university or laboratory in question would need to notify the government about patentable results within 60 days of realizing the patent potential, and indicate within 90 days the countries it would file patent in.

"One should not expect Indian universities to have an instant capacity to file and maintain patents, follow up on infringements, and negotiate technology transfers," adds Krishna Ravi Srinivas, an associate fellow at the New Delhi-based Research and Information System for Developing Countries. Still, he says, "a beginning has to be made somewhere, and Bayh–Dole is the best model we have".

Within the past few years, China, Brazil, Malaysia and South Africa have passed Bayh–Dole-type laws, hoping to capitalize on their research. But some experts argue that this approach may not work in developing countries.

Bhaven Sampat, an economist at Columbia University in New York City and co-author of a recent study on the subject (A. D. So *et al. PLoS Biol.* **6,** e262; 2008), argues that Bayh–Dole did not significantly affect the commercialization of US academic research. He cites 2006 data from the Association of University Technology Managers that suggest US universities, hospitals and research institutions derived \$1.85 billion from technology licensing — less than 5% of the money they received from federal, state and industry funders.

T. V. Padma

UN suspends leading carbon-offset firm

As international climate talks began last week in Poland, the United Nations (UN) suspended the work of the main company that validates carbon-offset projects in developing countries, sending shockwaves through the emissions-trading business.

Based in Oslo, Det Norske Veritas has in the past four years validated and certified almost half of the 1,200 projects approved under the 1997 Kyoto Protocol's Clean Development Mechanism (CDM). At its meeting on 28 November in Poznań, the CDM's executive board temporarily withdrew Det Norske Veritas's accreditation after a spot check carried out in early November at the firm's headquarters revealed serious flaws in project management.

The board did not specify which projects are affected, but cites problems with the company's internal auditing processes, and says that one of its staff members was verifying CDM projects without proper qualifications. As a result, "validation activities could not be demonstrated to be based on appropriate sectoral expertise", the board reports.

Det Norske Veritas is a risk-assessment and consulting company with about 8,000 employees in more than 100 countries. Its 2007 revenue was 8 billion Norwegian krone (US\$1.1 billion). It was the largest of 19 companies entitled to validate and certify projects proposed under the CDM, which aims to cut greenhouse-gas emissions by promoting climate-friendly energy technologies, such as wind or hydropower.

Certified emission-reduction credits from verified projects can be traded and sold on the emissions market, helping industrialized countries to meet their emissions-reduction targets under the Kyoto Protocol. However, only environmentally sustainable projects that would demonstrably not go ahead without additional revenue from sales of these credits are meant to be approved.


C. LARSON

Environmental groups have criticized the social impact of the Xiaoxi hydropower station.

PATHOGEN PROBLEMS
Budget woes for top UK biocontainment laboratory.
www.nature.com/news

Det Norske Veritas says that it is surprised by the board's strong reaction, and promised immediate action to regain its accreditation. "It is very unfortunate, but we don't want to start arguing about the decision," says Tore Høifødt, a senior vice-president and director of corporate communication for the company. "We feel that CDM, an emerging business, is still under development. We do have the required competencies, but we accept that we have to improve."

While suspended, Det Norske Veritas cannot propose new CDM projects to the UN for formal approval. Høifødt says that 20–30 projects currently in the process of validation are likely to be delayed, and that the company will not take on new projects for as long as the suspension is valid. It will continue to validate and verify ongoing projects.

Critics of CDM have long warned of possible conflicts of interest, pointing out that private companies such as Det Norske Veritas have a business interest in not deterring customers by being too critical about proposed projects.

The environmental campaign group International Rivers, based in Berkeley, California, for example, claims that firms such as Det Norske Veritas and the Munich-based TÜV-SÜD Group, the largest German verifying company, have been too lax in their assessment of the social and environmental impacts of hydropower projects in India and China.

The companies reject these accusations. "Our reviews, such as of the Xiaoxi hydropower project in China, are carried out in agreement with guidelines set up by the World Commission on Dams," says Thomas Oberst, a spokesman for TÜV-SÜD. The company has not yet had a spot check by the CDM executive board, but expects one very soon, he adds.

Since 2005, when the Kyoto Protocol came into force, CDM projects under way in 51 countries are thought to have saved some 250 million tonnes in greenhouse-gas emissions. The UN hopes that the scheme will help to abate almost 3 billion tonnes by the end of 2012, when the Kyoto Protocol is due to be replaced by a new treaty, currently being debated at Poznań.

But analysts say that ensuring reliable verification is a serious problem, and that the decision to suspend the largest player sends a powerful signal to others in the business.

"The UN has made clear it won't stand for lax validation," says Martin Kruska, a CDM expert with the First Climate Group, a carbon-trading consultancy near Frankfurt, Germany. "This is a warning not to be ignored." ■

Quirin Schiermeier

For a blow-by-blow report of the Poznań conference, see http://tinyurl.com/6edhzj.

Kansas wins race to host biodefence research centre

The US government is to put its $563-million National Bio and Agro-Defense Facility (NBAF) in Manhattan, Kansas. The bid from Kansas State University beat four university-linked applicants to replace the ageing Plum Island Animal Disease Center in Greenport, New York, as the leading national research centre on the most dangerous animal diseases.

But hiring the 300 or so trained researchers and support personnel needed for the centre and its biosafety level-4 (BSL-4) laboratory will be a stiff challenge — especially as the number of biocontainment facilities is rising nationwide.

D. MAYES, K-STATE PHOTOGRAPHIC SERVICES

Pat Roberts has supported Kansas State.

Sixteen months ago, Kansas State University finished building 21 BSL-3 labs as part of its $54-million Biosecurity Research Institute. But no research is under way there yet as the state's government has frozen recruitment, and only a few dozen staff members are preparing to be certified to work in the labs.

The concept behind the Biosecurity Research Institute was "build it and they would come", says its interim director Beth Montelone. University officials had hoped that biotech businesses or outside researchers would use the labs on a fee-for-service basis. Officials now hope that some Plum Island labs may move their research there before the NBAF is completed in 2014.

The other university-oriented bids to host the NBAF came from Georgia, Mississippi, North Carolina and Texas. The race had its roots in the anthrax terrorist attacks of 2001, after which the US government went on a spending spree to develop new biosecurity research complexes. It built one BSL-4 lab in Galveston, Texas, which was dedicated last month, and another is under construction in Boston, Massachusetts.

The Department of Homeland Security has also funded a series of 13 regional facilities of BSL-3 labs — in 2003, Kansas State University lost out to the University of Missouri for one of those. Last week, Kansas politicians began talking of their success in securing the NBAF before federal officials had announced their decision. Senator Pat Roberts (Republican, Kansas) boasted of his lobbying on his website. He was instrumental in earlier pushing through a federal $11-million appropriation for the Biosecurity Research Institute, which is housed in a hall named after him.

The inclusion in the NBAF shortlist of Mississippi — a site that ranked poorly in an initial review of 18 sites — raised some questions about the selection process. Asked if the final selection was political or scientific, Ron Trewyn, Kansas State University's vice-president for research, said: "I am very comfortable saying this was a scientific decision."

In its bid, the university highlighted that it has a veterinary college and BSL-3 labs that can accommodate animals up to 400 kilograms. It also flagged up the presence of the renowned husband-and-wife team of veterinarians Nancy and Jerry Jaax, and Juergen Richt, an expert in zoonoses who was hired last summer from a US Department of Agriculture facility in Iowa.

But even the best laid plans can't account for everything. In Texas, the new Galveston National Laboratory and its BSL-4 lab physically survived the devastation of Hurricane Ike in September — although its host institution, the University of Texas Medical Branch (UTMB), was forced to close temporarily with an estimated $700 million in damage. Nearly 4,000 people were laid off — mostly clinical personnel but also some 125 research-oriented staff, including the attorney who handles biosecurity issues for the Galveston lab. Last week, some of those laid off filed a lawsuit to get their jobs back.

Even after Ike, the Texas lab is expected to become operational next year; the Boston facility should follow soon after. To help cope with the rising demand, both UTMB and Kansas State University have recently launched programmes to train more workers for biosecurity labs. Whether it will be enough remains to be seen. ■

Rex Dalton

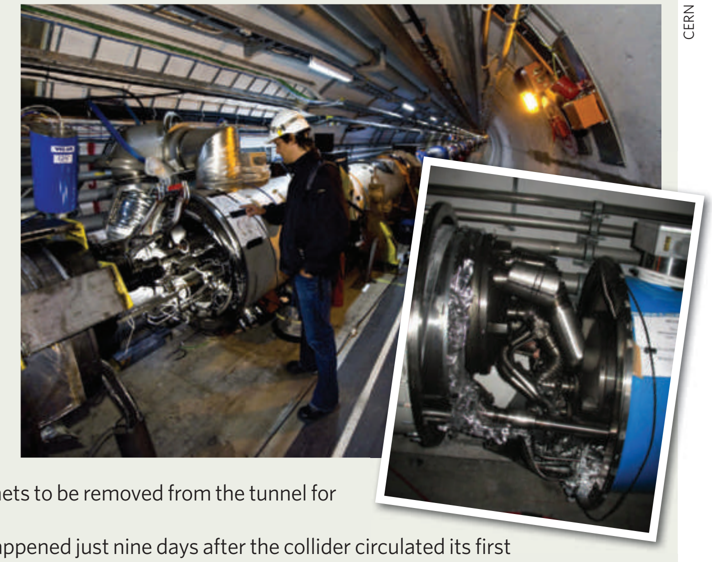
French research faces shake-up in reform plans

The board of France's National Centre for Scientific Research (CNRS) formally approved controversial reforms (see *Nature* 453, 573; 2008) on 27 November that would split the agency into nine semi-autonomous institutes, with a recommendation for a tenth in computing. The meeting required police protection from protesting researchers.

A draft of government proposals for a reform of French life sciences was also revealed last week by the newspaper *Libération* (see http://tinyurl.com/6ov4e8). It proposes the creation within two years of an overarching research council — the National Institute of Life Sciences — which would coordinate policy at the national biomedical agency INSERM, and health-related life sciences within the CNRS and other research agencies.

The report also suggests regrouping all life sciences within a single research agency, along the lines of a recent recommendation made by an invited foreign panel led by Elias Zerhouni, former head of the US National Institutes of Health.

Hadron collider faces long wait for repairs



CERN

The Large Hadron Collider (LHC), the world's most powerful particle accelerator, will not restart its beams before July 2009, according to a post-mortem report on an accident that shut it down on 19 September.

Repairs to the machine, located at CERN, the European particle-physics lab near Geneva, Switzerland, will cost up to 35 million Swiss francs (US$29 million) and will require 53 magnets to be removed from the tunnel for inspection and repair.

The accident, which happened just nine days after the collider circulated its first beams, sent thousands of amps of electricity coursing through the machine's refrigeration system, creating a hole (main picture) that released several tonnes of helium coolant, leading to further damage (inset). Technicians will install new diagnostics and additional safety systems around the machine's 27-kilometre ring to prevent a similar blowout happening.

For a longer version of this story, see http://tinyurl.com/4zp97a

Mars mission delayed by concerns over reliability

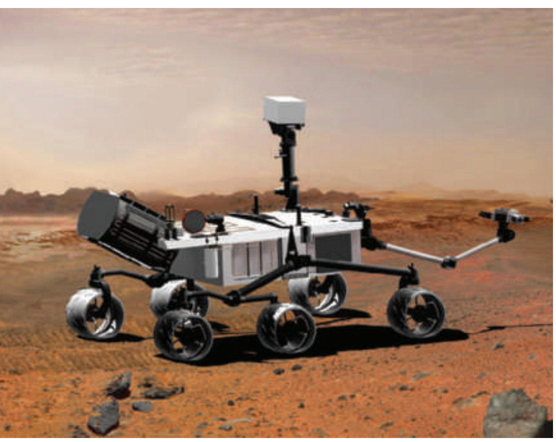
The launch of the Mars Science Laboratory, a one-tonne rover, has been delayed from October 2009 until late 2011.

NASA officials were concerned by the reliability of the motor units that drive both the rover's wheels and the joints of its robotic arm. The mission to look for signs of past life on Mars was slated to cost $1.6 billion when it was authorized in 2006, but now it is predicted to need between $2.2 billion and $2.3 billion to enable it to run until 2014. NASA expects to have to delay other missions to help cover the cost.

The increasing expense of Mars exploration prompted Ed Weiler, NASA's associate administrator for science, to announce that the agency will partner with the European Space Agency on future major missions.

For a longer version of this story, see http://tinyurl.com/57tahj

NASA/JPL-CALTECH



Unreliable? The Mars Science Laboratory.

Obama makes green choice for commerce department

US president-elect Barack Obama selected New Mexico governor Bill Richardson to head the commerce department on 3 December. The appointment highlights the importance of green jobs, energy and climate-change science within the agency.

The commerce department oversees economic policy as well as climate science, satellite operations and ocean fisheries regulation at the National Oceanic and Atmospheric Administration (NOAA), with a budget of $3.9-billion for fiscal year 2008. The National Institute of Standards and Technology also comes under the jurisdiction of the department.

Richardson was elected governor in 2002, and has made a broad push for the development of clean energy in New Mexico.

UK scientists promised £250 million for training

The UK Engineering and Physical Sciences Research Council is set to spend £250 million (US$370 million) on creating 44 centres to train 2,000 PhD students over the next five years. Expanding a small pilot scheme, the interdisciplinary centres for doctoral training will accept their first batch of students next October. Spread across 22 UK universities, they will focus on areas including climate change, sustainable energy, health-care technologies and nanotechnology. Seventeen of them will have strong ties to business: students will spend up to 75% of their time training with industrial partners.

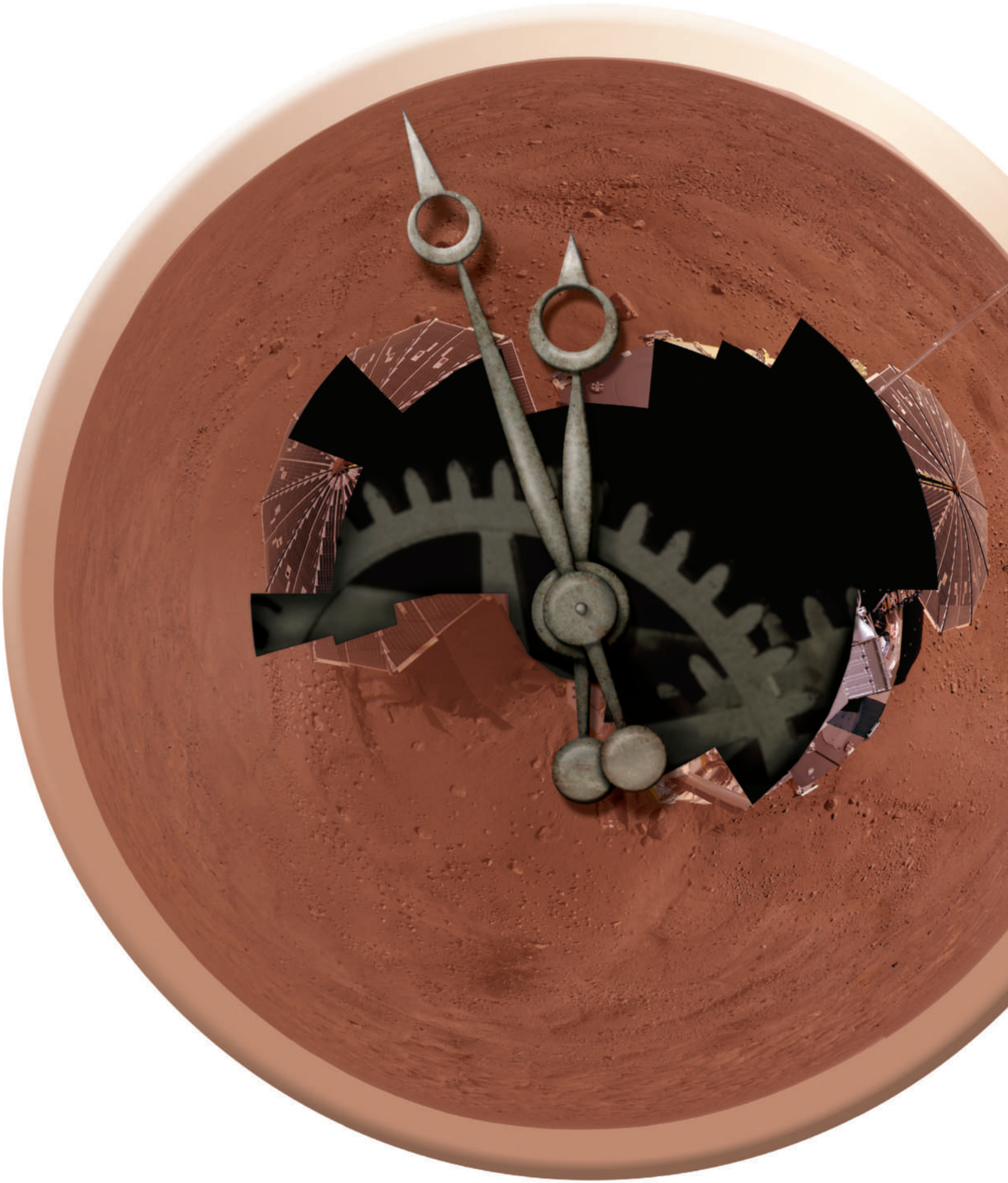
For a longer version of this story, see http://tinyurl.com/6xmedw

Institutions to disclose more commercial ties

An eminent US academic medical centre last week began publicly disclosing the financial interests of its 1,800 doctors and scientists. The website of the Cleveland Clinic in Ohio will list the collaborations with industry of every staff physician and scientist, including consulting relationships that pay $5,000 or more per year.

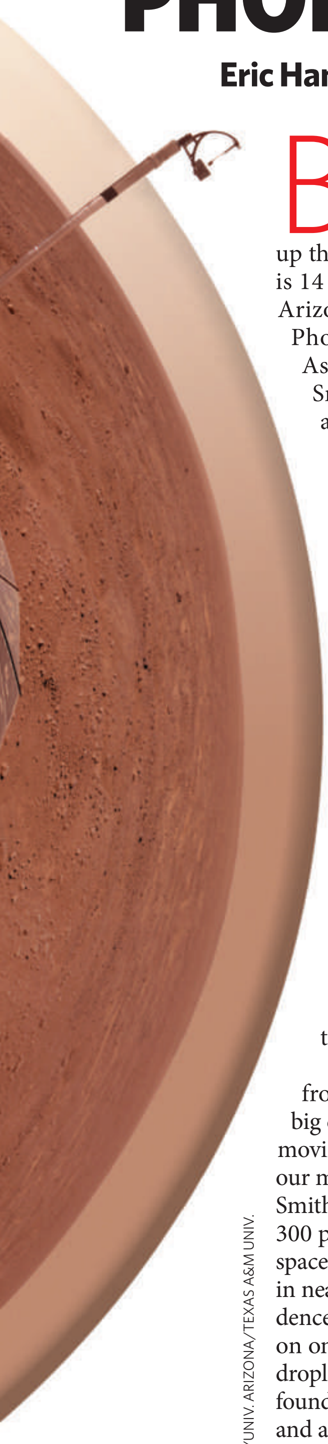
The medical school and health system of the University of Pennsylvania in Philadelphia plans to launch a similar disclosure system in spring 2009.

Other medical institutions are under growing pressure to follow suit. Senator Charles Grassley (Republican, Iowa) is aggressively investigating under-reporting of financial relationships by scientists funded by the National Institutes of Health. He has also introduced legislation that would oblige drug and device makers to make public all physician payments of $500 or more in cash or kind.



PHOENIX: A RACE AGAINST TIME

Eric Hand reports on the short life and hard times of the little Mars lander that sort-of-could.



NASA/JPL-CALTECH/UNIV. ARIZONA/TEXAS A&M UNIV.

Artwork based on a photomosaic that captures 360° of the flat landscape around Phoenix.

By the standards of Tucson, Arizona — let alone the northern plains of Mars — Ithaca, New York, is lush in any season. But winter is coming on. The autumn foliage is already past its prime as Peter Smith strides up the steep hill towards the Cornell University campus. It is 14 October, and Smith, a professor at the University of Arizona, is scheduled to give a talk on the status of NASA's Phoenix lander at the annual meeting of the American Astronomical Society's Division of Planetary Science. Smith is Phoenix's principal investigator, and the only academic ever to have overall responsibility for running a mission on the surface of Mars. He has decided to walk from his downtown hotel to the conference centre, and the unseasonably sunny weather is exacting its toll. A patch of sweat spreads from the centre of his bright orange golf shirt towards the burning, beady-eyed bird emblazoned on his breast: the Phoenix mission badge.

Smith is not the first speaker in the session on ongoing NASA missions; that is local hero Steve Squyres, the Cornell professor responsible for the science packages on Spirit and Opportunity, the rovers that have been trundling across the planet indefatigably since 2004. Squyres bounds up onto the stage in boots and blue jeans: "I've only got 20 minutes for this, so hang onto your seats." After a whirlwind tour of hills climbed, craters visited and geological features studied, Squyres finishes with the latest ambition for Opportunity: a 12-kilometre trek to a crater 22 kilometres across. It could take up to two years to complete.

Smith looks tired as he mounts the stage. He stands back from the podium and flashes a smile. "You know, there's a big difference between Steve and me," he says. "Steve's always moving. I stay in one place. I'm kinda a couch guy, ya know? So our missions are like that, too." But for all his self-deprecation, Smith talks with real pride about what his team of, at its peak, 300 people has done. It has delivered a comparatively cheap spacecraft to the surface of Mars, and set it down on a plain rich in near-surface ice. The analytical lab on board has found evidence of intriguing salts; Smith shows a picture of strange spots on one of the lander's legs that might, conceivably, be water droplets (see 'Strange brew', page 692). Another instrument has found evidence for carbonates, formed in the presence of water, and a weak signal that, Smith says, might just be due to organic molecules — something never detected before on Mars. So much to relish and pursue — if only there were more time.

But there isn't. Whereas Squyre's rovers have had their mission stretched from its original 90 sols — the term used for a Martian day — to 1,700 and counting, nothing like that is possible for Phoenix. The plain it sits on is far to the north of the rovers, and the winter that is swiftly coming on is harsh enough to freeze the thin atmosphere onto Phoenix's body. The scientists running the various instruments are jockeying for a share of the ever-lower power levels as the days get shorter and the sun sinks lower; the end is in sight. The Cornell talk is on the mission's 138th sol. Smith knows that well before sol 200 the mission will be over, the lander dead on its darkling plain.

Reversal of fortune

The mission has taken its toll on Smith, normally a gregarious, happy-go-lucky man. During the talk, he charms the audience with his humour. But over lunch he is uncharacteristically downbeat, even testy. "It's not quite what I thought it would be at the beginning," he says, picking at his food. "We're right there, next to a soil that has all these wonderful secrets locked into it. We've got the right instruments, we've got the right people involved, everything is perfect. And we can't quite get those bits of scientific knowledge out of our instruments. It's a frustration, I tell you. We're going right down to the wire."

> **"The mission's not quite what I thought it would be at the beginning."**
> — Peter Smith

Phoenix was a child of misfortune. After the success of Mars Pathfinder in 1997, the first landing on Mars since those of the Vikings in 1976, NASA's Mars programme managers had taken up then-Administrator Dan Goldin's mantra of "faster, better, cheaper", trying to deliver two small, innovative spacecraft to the planet every two years. But in September 1999, Mars Climate Orbiter burned up in the planet's atmosphere because no-one had noticed a confusion between metric and imperial units in the navigation commands. Three months later Mars Polar Lander (MPL) was lost as it made its way to a site near the planet's South Pole, probably because its thrusters shut off while it was still 40 metres up in the air.

Taking no more chances, in 2000 NASA cancelled a lander planned for 2001 that would have used the same design as MPL. And a proposed orbiter, the Mars 2003 Surveyor, lost its launch opportunity to Squyres' rovers. This put Smith in dire straits. Smith had run the main camera on Mars Pathfinder, and had MPL survived he would have done the same on that mission. He had also been working on providing and running cameras for both the cancelled missions, work now curtailed. "I had no job. I lost 35 employees. I had no mission." Smith went to work for University of Arizona colleague Alfred McEwen, who was managing the Hi-Rise camera for the Mars Reconnaissance Orbiter, due to fly in 2005.

But faster, better, cheaper was not quite dead. In 2002, NASA opened up a competition for a new 'Mars Scout' line of low-cost missions, modelled after its Discovery missions to the rest of the Solar System. Smith got back into the fray. He was the cameraman on six of the Scout proposals — and the principal investigator on a proposal of his own. "No matter who won," he says, "I would win too."

Smith's proposal, Phoenix, was an ingenious one. It would use hardware from the abandoned 2001 lander to run a mission

A. POULSON

L. K. HO/AP

Sol 0: Principal investigator Peter Smith celebrates Phoenix's safe touchdown, as depicted on the mural outside the mission's science operations centre.

similar to the ill-fated MPL — but flipped from one pole to the other. Bill Boynton, a burly, balding colleague of Smith's at the University of Arizona, and also a veteran of MPL, was in charge of a γ-ray spectrometer on Mars Odyssey, an orbiter launched in 2001. He was finding evidence of broad haloes of near-surface hydrogen extending out from the planet's polar ice caps. The implication was that there was water ice in the plains, and that it was close enough to the surface to be studied with little more than a trowel. The northern plains seemed to be as icy as MPL's target site in the south, but with the advantages of being flat, mostly boulder-free and at much lower elevation. A spacecraft landing there would have considerably more atmosphere to slow it down before reaching the surface.

The Phoenix proposal suggested a landing on those northern plains with two specific goals: to study the history of water in the Martian arctic, and to assess the biological potential of the boundary between the ice and soil, in search of evidence of an environment that might be habitable. In 2003, to the surprise of many, Phoenix won the competition to become the first Mars Scout, beating a range of less risky missions. As on the Discovery missions, Smith would be in charge of everything. A project manager at the Jet Propulsion Laboratory (JPL) in Pasadena, California, would supervise the contracts and construction of the mission. But Smith's out-of-the-way university would be the operational home.

So beautiful, so new

Smith's science operations centre is in a residential neighbourhood a few kilometres from the university campus in Tucson. It is a low-slung, stucco building amid the dun-coloured homes, anonymous but for a bright, almost garish mural showing Phoenix's descent to Mars. On 25 May — sol zero — hundreds of scientists, engineers and their families gathered there to munch on picnic food and watch the tracking information relayed to Earth via Odyssey and the Mars Reconnaissance Orbiter as Phoenix's interplanetary voyage came to an end.

In what the team was calling the 'seven minutes of terror', Phoenix plunged through the thin air. The blackened heat shield was jettisoned. The parachute opened. Mission controllers called out altitudes. Twelve retro-rocket thrusters began a rapid fire dance of high-pressure hydrazine. After a final twist to orient itself to maximize the amount of sunlight its solar panels would absorb, Phoenix touched down. It was the first landing on Mars since the Vikings to use rockets, not air bags, to cushion the descent, and in the imaginations of the scientists and engineers back on Earth it had done so flawlessly. In Tucson, people cheered and clapped and cried. "The landing

Strange brew

"People always joked about finding oil on Mars," says Michael Hecht of the Jet Propulsion Laboratory in Pasadena, California. "What more could you ask for than to take a shovel-full of soil, rinse it with water, and have pure rocket fuel?" In terms of human exploration, says Hecht, "we actually found gold!".

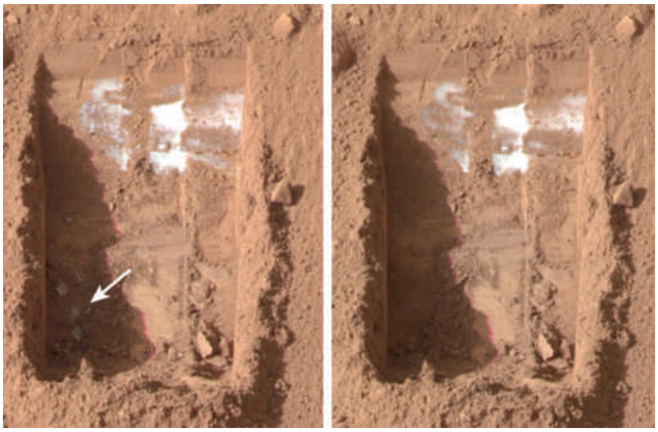
In fact, Hecht and his colleagues on Phoenix's wet-chemistry experiment had found perchlorate ions. Perchlorate is a powerful oxidizing agent — that's why, in the form of ammonium perchlorate, it is used in rockets. And the idea that the Martian soil might contain oxidizers has been around since the days of the Viking landers as an explanation for the perplexing lack of organic molecules (see *Nature* **448,** 742–744; 2007), with peroxides favoured for the task. When Hecht and his colleagues found perchlorate levels of a few per cent in the slurries of soil and terrestrial water they made in Phoenix's little chemistry beakers, they immediately saw they had come up with an interesting alternative.

But Hecht and his team got even more excited about some of the other properties of perchlorate salts. They have a great avidity for water and a low-temperature stability that makes them much less toxic than peroxides. Perchlorate salts can also, at high enough concentrations, make a pretty good antifreeze. These properties had implications for the possible habitability of the Martian subsurface — and also for its study. The absorptive tendencies might explain the stickiness that made the mission's soil samples so hard to handle.

The stability means that, unlike peroxides, perchlorates in the Martian soil would not necessarily be poisonous. They might even be food. On Earth, water contaminated with perchlorate is treated with bioremediation — that is, by getting microbes to eat it up. And the antifreeze effect could, conceivably, allow water to stay liquid in the comparatively warm conditions of a high-latitude Martian summer. If you want

was probably the biggest high of the whole mission," says Smith. "We were told over and over and over again to expect it to fail."

The influx of data was, at first, as smooth and splendid as the landing had been. As the scientists adapted their bodies to sols instead of days — an extra 40 minutes every 24 hours leads to something like permanent jet lag — Phoenix's main camera, mounted two metres above the plain, started on a panorama. Early shots revealed cracks in the soil that marked out polygons of different sizes. The cracks hinted at cycles of warming and cooling that had changed over time — in sync, perhaps, with oscillations in the tilt and orbit of Mars. The lander's robot arm reached out and scratched at the ground, and on sol 20 the camera saw some pale nuggets at the bottom of a trench a few centimetres deep. Four sols later they were gone; it was ice that had sublimed from solid to vapour after being exposed to the balmy warmth of the polar summer: –31 °C that day, according to the lander's meteorological station.



NASA/JPL-CALTECH/UNIV. ARIZONA/TEXAS A&M UNIV.

Sol 20: Phoenix spots possible chunks of ice (arrow) in the Dodo-Goldilocks trenches. By sol 24 they are gone (right).

The pictures were meant to be just an appetizer; the main course was to be soil samples fed into Phoenix's eight tiny bellies, each of them an oven that could bake its contents to 1,000 °C. Phoenix's nose, a mass spectrometer, would sniff the gases given off, detecting compounds down to concentrations as low as 10 parts per billion. The whole ovens-plus-spectrograph system was the Thermal and Evolved Gas Analyzer (TEGA), and it was crucial to the mission's goal of seeking out organic compounds in the soil. Viking had failed to find such compounds; but TEGA had hotter ovens and a landing site where, some argued, the iciness of the soil would preserve organic compounds well (see *Nature* **448,** 742–744; 2008). Such compounds might not be evidence of life — various apparently lifeless sites in the Solar System have complex carbon chemistry — but they would speak to Phoenix's goal of assessing habitability.

TEGA, based on a similar instrument that had flown on MPL, was run by Boynton. The man who was detecting ice from orbit with one spectrometer would also melt it down on the Martian ground to sniff its isotopes with another. TEGA was at the heart of Phoenix's planned science, Boynton says. Its results were the most eagerly anticipated, and its care and feeding took up a great deal of the team's time. Impressively, the workhorse had been put together for just $13.6 million by a dozen university scientists and technicians. But TEGA was temperamental from the moment it was put to work.

By this distant northern sea

In April 2007, a few days before TEGA was shipped from Tucson to Lockheed Martin in Denver, where it would be bolted onto the spacecraft, a short circuit was found in a filament inside the mass spectrometer. It was not a fatal flaw; there was time to clear up the problem, and the instrument had a second filament as backup. On sol 4, though, the backup filament was found to have a short circuit, leaving TEGA without a safety net and heightening the stress on Boynton's team. Towards the end of June, on sol 25, another short circuit was

to fuel speculation, rather than rockets, the possibility of liquid water and microbe food on Mars is about as good as it gets.

Nilton Renno, a member of Phoenix's science team from the University of Michigan at Ann Arbor, became fascinated by that possibility. Early in the mission, Renno suggested that little specks on the lander's legs were splashes of salty mud kicked up by its retro-rockets during the landing. However, when the camera returned to check on the legs weeks later, the images suggested something extraordinary: the specks might be growing.

Renno now says that the specks — which he claims are not only growing but also merging and moving — had become briny water droplets. The lander legs, he says, are about 20 °C warmer than the surface because of the spacecraft's heaters. This, combined with the effects of perchlorate, allowed droplets to form. If liquid water exists on the lander leg in just a few randomly spattered spots, he speculates, what about in the soil itself? On Earth, small layers of briny water exist within massive chunks of sea ice. Why couldn't the same be true in the boundaries between ice and soil on Mars where salts might be expected to accumulate? "I'm not saying that there's liquid water everywhere on Mars. I'm just saying that these results suggest that it is possible," says Renno.

The idea has been hotly debated within the team. Renno says he will submit a paper to the *Journal of Geophysical Research* that describes the droplets and asserts that briny water is "common". He says he has two dozen co-authors from the science team. But there are some notable absences on the list, including Hecht and Ray Arvidson, a senior mission scientist from Washington University in St Louis. "My impression is that Nilton is running with the most spectacular hypotheses," says Arvidson.

Principal investigator Peter Smith, who has not yet decided whether to stay on the paper as a co-author, says the 'barnacles' might have been created chemically by the lander's retro-rockets. "Maybe it's telling us more about thruster interaction with the surface than it's telling us about Mars."

Hecht agrees with Renno that the specks on the lander leg are probably made of H_2O — but the frozen kind. He says that the lander legs, mostly in shadow, are likely to be colder, not warmer, and that the shifting of shadows can cause the temperature variations needed to explain the growth and movement of the ice. "We would all love to find liquid water dripping around Mars," says Hecht. "I think it's an extraordinary claim. I say, 'do you have the extraordinary proof?'".

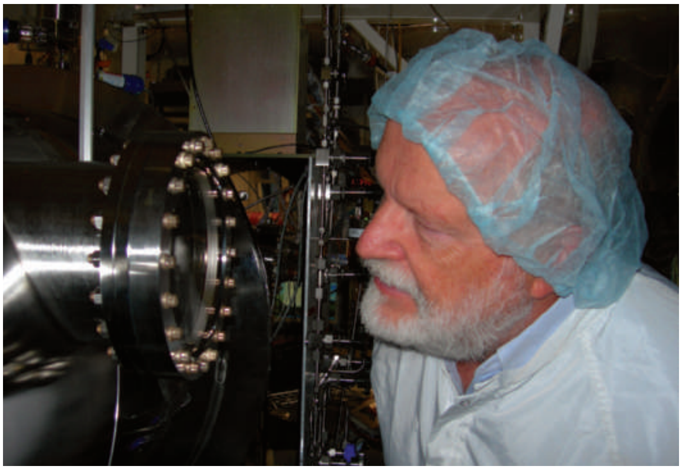
The mission's last scoopful

NASA/JPL-CALTECH/UNIV. ARIZONA/MAX PLANCK INST.

Sol 8: Specks on the lander's legs could turn out to be briny water.

might have gone some way to providing that, with a sample from comparatively deep within a trough between two 'polygons' — a good spot for perchlorate to accumulate. Unfortunately the sample was so sticky it couldn't be dislodged into the chemistry beaker. Perchlorates giveth, perchlorates taketh away? **E.H.**

H. ENOS



Bill Boynton oversees development work on Phoenix.

discovered, this time in one of the oven units, probably caused by the shaking the unit was subjected to as the spacecraft operators tried get some surprisingly sticky soil through a grating. Again, the short wasn't fatal — it affected only one of the ovens — but it added to the nervousness both in Tucson and at NASA headquarters.

On missions led by principal investigators, such as the Discoveries and Scouts, NASA is supposed to defer to the scientist in charge on all matters of scientific operation. But Phoenix was high profile and some of its instruments a little erratic. At headquarters, everyone from Administrator Michael Griffin down was involved in daily reviews of the mission, says Doug McCuistion, Mars exploration programme chief at NASA. At the end of June, word came down that the Phoenix team was to treat its next TEGA sample as its last, and to go after a sample of rock-hard ice before it did anything else. The Tucson team had lost its autonomy.

"We stepped in, I'll be honest," says McCuistion. Boynton — a bit of a bulldog when it came to keeping control over his instrument — acknowledges the logic: "NASA was really afraid … that if we never got the ice it would be embarrassing." But he and Smith still resent the way that the mission was taken over. "That's not the way you do these things," says Smith. "That's why we were pushed at the end."

It took the team days to figure out how to get the scraps of ice, shredded with a rasp, into the arm's scoop. Then, when the scoop was turned over above one of TEGA's ovens, the ice refused to fall out. Even when the team figured out how to sprinkle some out, it was faced with what would become the most nagging of the mission's problems: the doors to the ovens only opened partway. Much of what came out of the scoop didn't make it into the ovens.

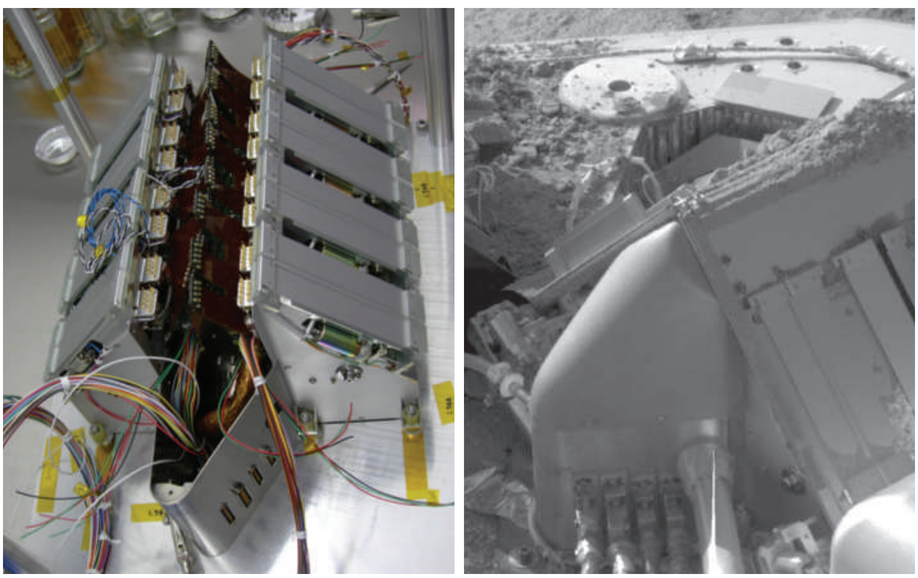
What made the glitch most maddening was that the mistake had been caught before launch. One of the differences between the TEGA on Phoenix and that on MPL was a thin retractable cover to keep the instruments from being contaminated by any stowaway microbes from Earth. Boynton and his team had noticed, on a test version of TEGA, that the brackets at the bottom of this cover were just a hair's width too big, and as a result obstructed the doors. They sent revised designs for the cover to the manufacturer, Honeybee Robotics of New York. New parts were delivered and installed. But Honeybee had made the new parts using the original flawed designs — and nobody in Tucson checked them. "They should've caught it and we should've caught it, but neither of us did," says Boynton, ruefully.

Boynton says that the problems with TEGA weren't lost capabilities, but lost time: "the clock was running against us". After three unsuccessful weeks attempting to get some pure ice into an oven, the NASA directive was relaxed, and the Phoenix team went for its preferred option: scrapings of icy soil. There was enough ice in the soil for TEGA to confirm that it was water, but not enough to measure its isotopic make-up — a measurement that could have provided insights into the history of the planet's water and atmosphere. In August and September, TEGA went on to cook up several more shallow soil samples, and found strong signals for calcium carbonate, which is typically found precipitated out of water. More intriguing to Boynton was a low-temperature signal in all the samples — the same signal that Smith had hinted at in Ithaca. It probably came from a different type of carbonate, but it could have been the trace of an organic molecule.

The breath of the night wind

On sol 153 — a few days before Halloween, and a few weeks after Smith gave his talk at Cornell — Boynton celebrates his 64th birthday and convenes the final TEGA planning session in his office on campus. The mission scientists have long since vacated the science operation centre; only Smith and some support staff remain there. Boynton has a bandage on his left hand where a mishap with a coffee machine has left him with a third-degree burn; TEGA hasn't been much kinder. The previous evening Boynton had heard from mission engineers that Phoenix had entered a 'safe' mode — the "do no harm" response programmed into spacecraft as a way to deal with unexpected circumstances — and that commands to Phoenix to shut down its heaters, an attempt to conserve power, had not gone through. But there is still a chance that the TEGA team will get the power, and time, for one last experiment. With his good hand, Boynton wipes crumbs of celebratory chocolate and pumpkin cake from the table and he and his engineers sit down to go over blocks of programming code to be radioed up to Phoenix. They have done this hundreds of times before in the lifetime of the mission. Sunlight streams

UNIV. ARIZONA; NASA/JPL-CALTECH/UNIV. ARIZONA/MAX PLANCK INST.

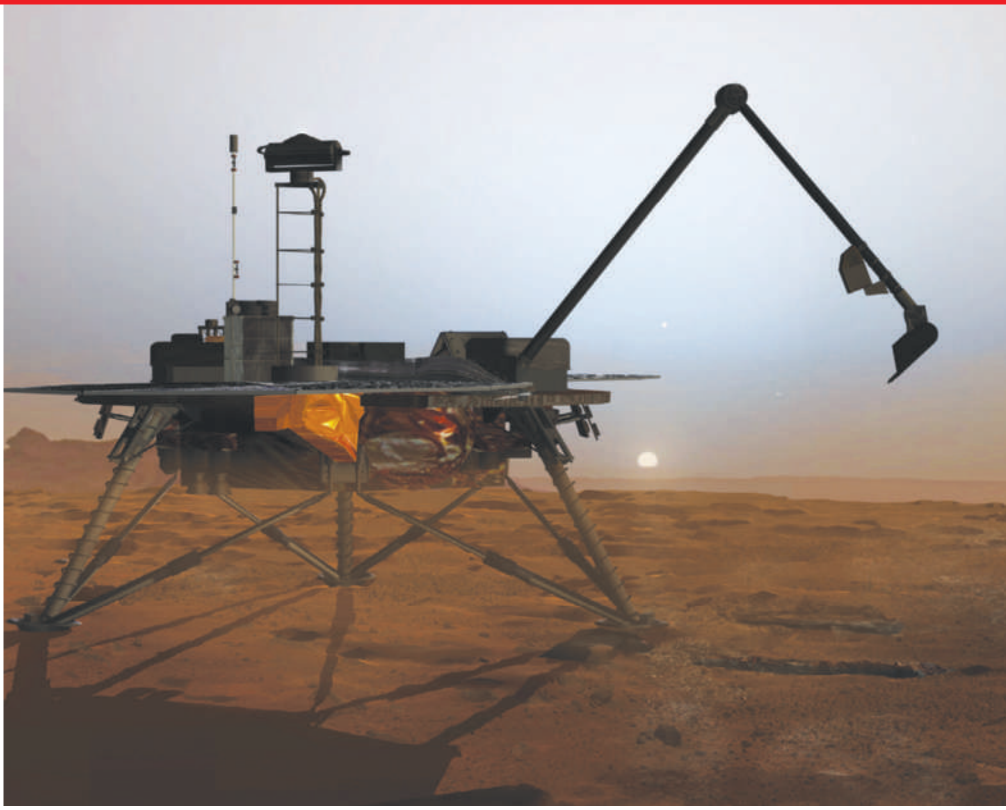


The Thermal and Evolved Gas Analyzer looked fine in the lab (left), but on Mars, soil samples didn't always make it inside.

in through a large glass window, filled nearly to the edges of the frame by the crags of Mount Lemmon, which looms over Tucson. "This is the last block I'm ever going to have to write," says one of the engineers.

They pour over their packet of code — instructions to some valves to open and some to shut — looking for errors. The goal is to suck an atmospheric sample into TEGA and draw out all the carbon dioxide, leaving proportionally higher concentrations of trace gases. The mass spectrometer would then measure the isotopes of argon and other noble gases, which in turn would offer up information about the history of Mars's atmosphere.

They know their best-laid plans may not come to pass, just as they have not in the past. They are ready for the end, which will bring both relief and grief. "It's really has been a lot more emotional than I expected it to be. We wanted it to do a little bit more," says Heather Enos, the TEGA instrument manager. "But isn't that always human nature?".



C. WASTE/JPL

Sol 157: Phoenix's final sunset, as imagined by an artist on Earth.

In many ways, the trials and tribulations of TEGA are representative of those of the entire mission. At $428 million, Phoenix was a 'low-cost' mission that took on significant risk. According to Boynton, there is no way that one of NASA's traditional operation centres, such as JPL, could ever have operated it for less than half a billion dollars, or built an instrument like TEGA for $13.6 million. "JPL would just say you can't do it, you gotta do it for this much or not at all. We figured there was a shot we could do it. But we did cut corners." A similar, but much fancier, instrument on JPL's vastly overbudget $2-billion Mars Science Laboratory, the launch of which was last week put off for two years to 2011, has cost about $80 million. It is being assembled by 50 scientists and engineers.

Gentry Lee, JPL's chief engineer for Solar System exploration, says that Phoenix will be remembered for getting "remarkable bang for the buck". The lander performed science on almost every sol of its 157-sol life. It baked samples in five of eight of its ovens, and used three of its four wet chemistry beakers; it never tested the ice, but it tested icy soil. It took more than 25,000 pictures. Its atomic force microscope saw plate-like particles 100 times smaller than those spotted in the best resolution images from the rovers' microscopic imagers. Its main camera saw snow falling from overhead clouds, and frost gathering on the ground.

Boynton says that they didn't reach the optimum point on the science per dollar curve, and he wishes, especially, that they had more money for post-mission analysis, given that the scientists had so little time to work with data during the mission. He is still working on Earth-based controls that might help to decipher that enigmatic low-temperature, probably carbonate signal from TEGA. But he doesn't regret the way things turned out one bit. They stayed within their cost caps. And they delivered. Maybe not to the doorstep, but at least to the front yard. "You can't expect the toast to always fall butter-side up."

> "There's some relief when it's over. On the other hand, it seems over too soon."
> — Peter Smith

Like a land of dreams

The next night, when Phoenix finally takes leave of this world, Peter Smith sleeps through its slipping away. At seven the following morning, he answers the door to his home in NASA-labelled athletic shorts, having just woken up to an e-mail with the subject 'Not so good news'. "You picked a heckuva day to come," he says. The sun, still low, has yet to bake off the desert's night-time chill.

Despite efforts to turn off power-hungry heaters, the lander, with its batteries drained, has gone beyond its safe mode to 'Lazarus mode', an autonomous state in which it tries to operate fixed programs but no longer takes new commands on board. Engineers will monitor weak signals for three more days but never regain control. The science mission has ended.

But the time to grieve has not yet come, because the dog must be walked. Smith puts on his running shoes and grabs the leash. Happy, his exuberant one-year-old mutt — "some sled dog," Smith mutters, still bleary — is already bouncing. Happy charges past the barrel cactus in the front yard and Smith, tugged along, wrestles for the words that describe the end that he knew would come. "There's some relief when it's over. On the other hand, it seems over too soon."

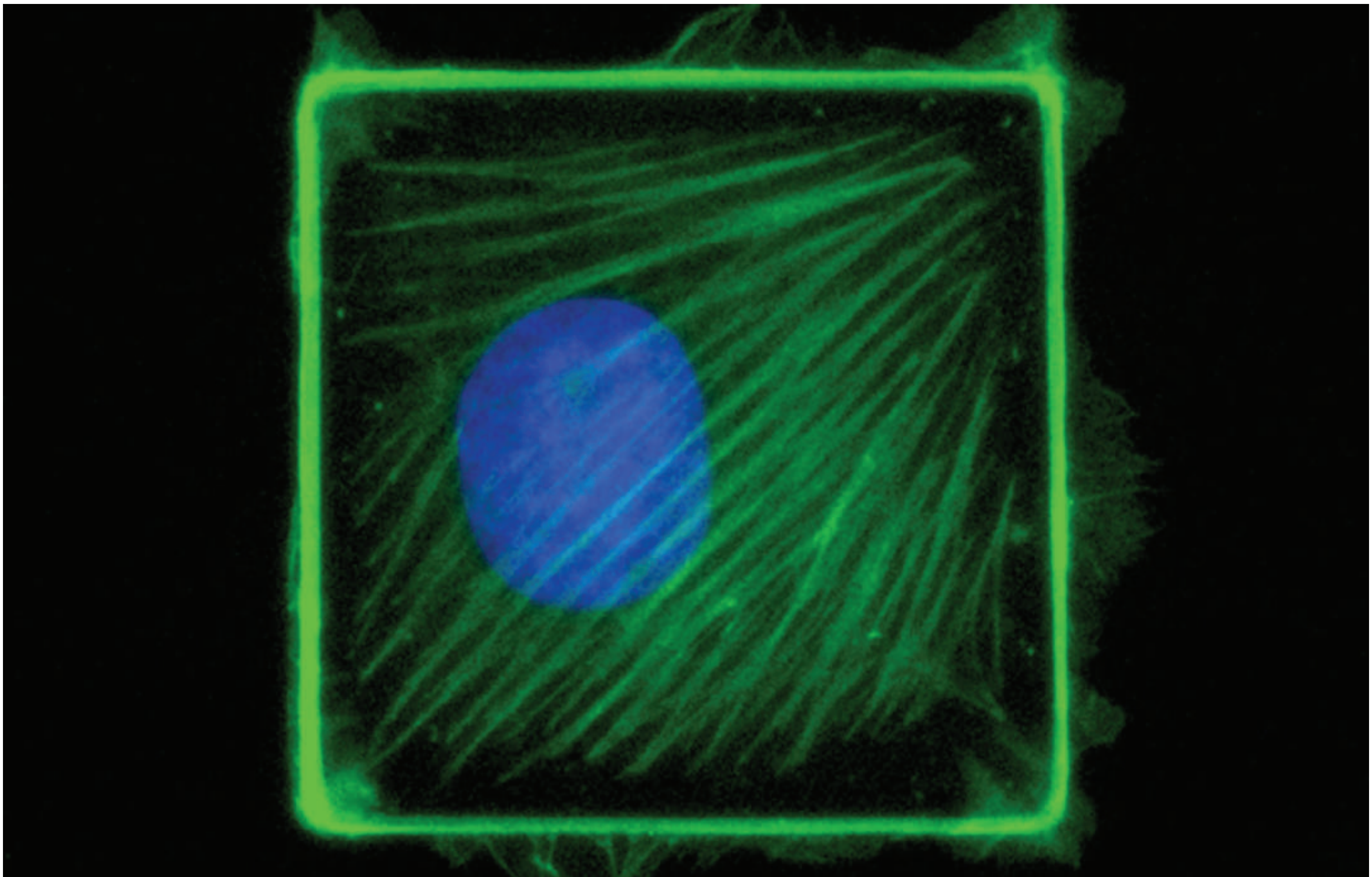
Lee, who has seen generations and missions come and go, says that there is a place for Mars missions of all styles and sizes. Big, expensive flagships for the jobs that cannot fail; smaller, cheaper missions for the ones that can; and, occasionally, medium-sized ones that take dead parts and make them live again. Phoenix has shed light on the results of the Viking experiments, and it is already influencing the way that scientists practice with the scoop on the Mars Science Laboratory. Mars missions are mutually indebted in a way missions to more seldom visited places cannot be — it is part of the luxury of a destination less than a year's rocket ride away. An exploration programme committed to staggered missions, built on what came before, provides some solace in the face of disappointment. But it's hard to deny the pathos of a mission that goes from landing to ending in five months, with many hopes unrealized.

Tucson is getting hotter by the minute. In a desert 374 million kilometres away, Phoenix is chilled to the bone, waiting for the solid CO_2 that will soon freeze from the sky to entomb it. Peter Smith, 61, will soon be out of his job as a principal investigator, and he will never run a spacecraft again. Happy noses in the dirt, sniffing for something that's probably organic. ■

Eric Hand writes on physical sciences from *Nature*'s Washington DC office.

See Editorial, page 675.

Correction
The News Feature 'Phoenix: a race against time' (*Nature* **456,** 690–695, 2008) suggested that Honeybee Robotics failed to follow revised designs that would have corrected a flaw in a part of the NASA Phoenix spacecraft. In fact, the revisions were not clearly marked for action. A subsequent investigation by mission managers at NASA's Jet Propulsion Laboratory found that Honeybee was not at fault. See Correspondence, page 26.



C. CHEN & D. INGBER

STRETCHING THE IMAGINATION

Squash them, pinch them, twist them, pull them — cells react to physical forces, finds **Claire Ainsworth**.

The lecturer flattened a tangle of sticks and elastic on to the desk. He let go, and the structure pinged back into shape. It was a demonstration of 'tensegrity', a term constructed by the engineer Buckminster Fuller for situations in which push and pull have a 'win–win relationship' with each other. Fuller mashed the word together — from the components tensional integrity — to describe the way sculptor Kenneth Snelson used taut wires and stiff poles to make strong yet flexible monuments.

Among the students, Donald Ingber could see the ingenious engineering in the sculptures — but he also saw biology. Ingber was an undergraduate in molecular biophysics at Yale University; the course in three-dimensional design had seemed apt. But what he saw there changed the course of his professional life. At that time, in the late 1970s, researchers were publishing the first scientific papers describing how cells are propped up by an internal scaffolding, or cytoskeleton. "I immediately thought: 'Oh, so cells must be tensegrity structures'," Ingber says. On returning to the lab, he eagerly explained this idea to one of the postdocs, who was less than impressed. "He told me: 'Just never mention that again'," Ingber recalls.

Then, as now, most cell biologists had little time for architecture and engineering. When they want to understand why a cell behaves the way that it does, they try to identify the genes, proteins and signalling molecules that are thought to exert control. But to Ingber there was an obvious gap between the dramatic events that mould a developing embryo and the molecular explanations that were given for them in his developmental-biology class. "What I saw before my eyes was something that was incredibly physical, mechanical in nature: twisting, bending, folding," he says, "and then I got into cell biology, and it was all chemical."

Thirty years on, work from Ingber's group and many others has started to convince cell biologists to embrace the missing physics. Their findings are remarkable. Pull a stem cell in one way and it starts developing as a brain cell; stretch it in another, and a bone cell is its more likely fate. Change the mechanical stresses on cancer cells and they can start to behave more like healthy ones. Among this work's implications, few are more important than the consequences for cell therapy and tissue engineering, in which researchers hope to use new cells to repair damaged organs. If these cells encounter the 'wrong' kinds of mechanical stresses, they could conceivably end up doing more harm than good.

The discoveries are giving biologists a fresh appreciation of the body's physical nature. Hearts pump, muscles stretch, blood surges, feet pound. And on the microscopic scale, fluids flow and cells jostle with their neighbours. When Ingber, now at Harvard University, talks about his ideas today, he doesn't get quite the same frosty reception that he once did. "There's no doubt," he says, "in the past five years it has exploded."

Popular mechanics

Even in the 1970s, the ideas were not entirely new. The importance of mechanical forces was appreciated by embryologists in the 1800s and early 1900s, long before they had signalling proteins and chemical gradients to play with. Swiss biologist Wilhelm His, for example, experimented with metal, clay and rubber to try to mimic events in development, such as

how the future brain starts forming as a roll of tissue on the back of a mammalian embryo. "To think that heredity will build organic beings without mechanical means is a piece of unscientific mysticism," he wrote in 1888 (ref. 1).

The molecular biology revolution of the 1960s onwards pushed much of this aside, as researchers focused on genes and proteins. But the mechanical ideas never fell completely out of favour. Physiologists know that astronauts' bones get thinner when they escape gravity, and that hefting weights inflicts physical damage on muscle cells that stimulates them to grow. But it was widely felt that the role of mechanical stress would be limited to these and other cell types that needed it in order to function normally.

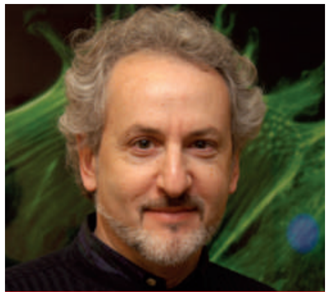
Then, in 1978, Judah Folkman and Anne Moscona at Harvard Medical School published one of the first studies to experimentally stretch mammalian cells — in this case, cells extracted from cow's blood vessels and other tissues. They coated plastic culture dishes with various concentrations of a sticky polymer and grew the cells on top. The stickier the substrate, the flatter the cells stretched out, and the more the cells stretched out, the more they divided. It was considered a landmark paper — one of the first to show that cell shape influences growth[2]. And it had a "profound effect" on Ingber when he read it that year. "It resonated with my sense that physicality was critical to developmental control," he says. Ingber later contacted Folkman and eventually went on to complete a postdoc in his lab.

Researchers now know that almost all human cells test the mechanical properties of their microenvironment in the body, and use it to adjust their growth. This is determined by the extracellular matrix, a lattice of proteins and other molecules to which cells in solid tissues anchor themselves like a horde of old canvas tents staked out at a rock festival. Instead of poles and ropes, cells have their internal cytoskeleton. This includes a mesh of fibres made up of actin protein that lines the cell's membrane, plus tough 'actomyosin bundles' in which actin combines with the protein myosin II. The tent pegs in this case are proteins called integrins that span the cell membrane, gripping the actomyosin filaments inside the cell and the extracellular matrix on the outside.

In 2006, Ingber's team used a femtosecond laser to cut the actomyosin filaments and found that they immediately retract, revealing that they are under stress just like the tensed wires in Snelson's sculptures[3]. The role of the poles seems to fall to another part of the cytoskeleton, the microtubules, which buckle under severe stress.

Bioengineers, meanwhile, have attempted to measure the elasticity of various tissues in the body. The simplest way involves hanging a weight from a hunk of tissue and measuring how much it stretches. More sophisticated methods include atomic force microscopy, which uses a cantilevered tip to lightly prod cells and measure their springiness.

Last year, bioengineer Paul Janmey and his colleagues at the University of Pennsylvania, Philadelphia, used atomic force microscopy to show that connective tissue cells called fibroblasts can tune their internal stiffness to match that of the substrate on which they are growing[4]. "That was actually quite a surprise," says Janmey. He thinks that this type of surveillance could help cells respond to change in the tissue around them. Imagine a cell that is normally pegged down on all sides in the skin — but whose moorings are weakened when the skin is cut. The physical change could be felt just as fast, if not faster, than chemical signals released by the wound. "The cell already knows that mechanically, something has gone haywire and can respond to that," says Janmey.



K. JOHNSON

"What I saw before my eyes was something that was incredibly physical, mechanical in nature."
— Donald Ingber

Ingber says that this idea is already being used in hospitals to promote healing, an idea he is working on in collaboration with Dennis Orgill, a plastic surgeon at Harvard's Brigham and Women's Hospital. Orgill seals a sponge-like device over hard-to-heal wounds such as those left by deep surgery, or diabetic foot ulcers. He then attaches this to an oscillating suction pump. The suction stretches the cells so that they divide, form new blood vessels and regenerate skin tissue, thus healing the wound[5] — the same type of behaviour that Folkman observed in his cultured cells 30 years ago. Ingber says the system works better (and is far cheaper) than artificial skin or synthetic growth factors.

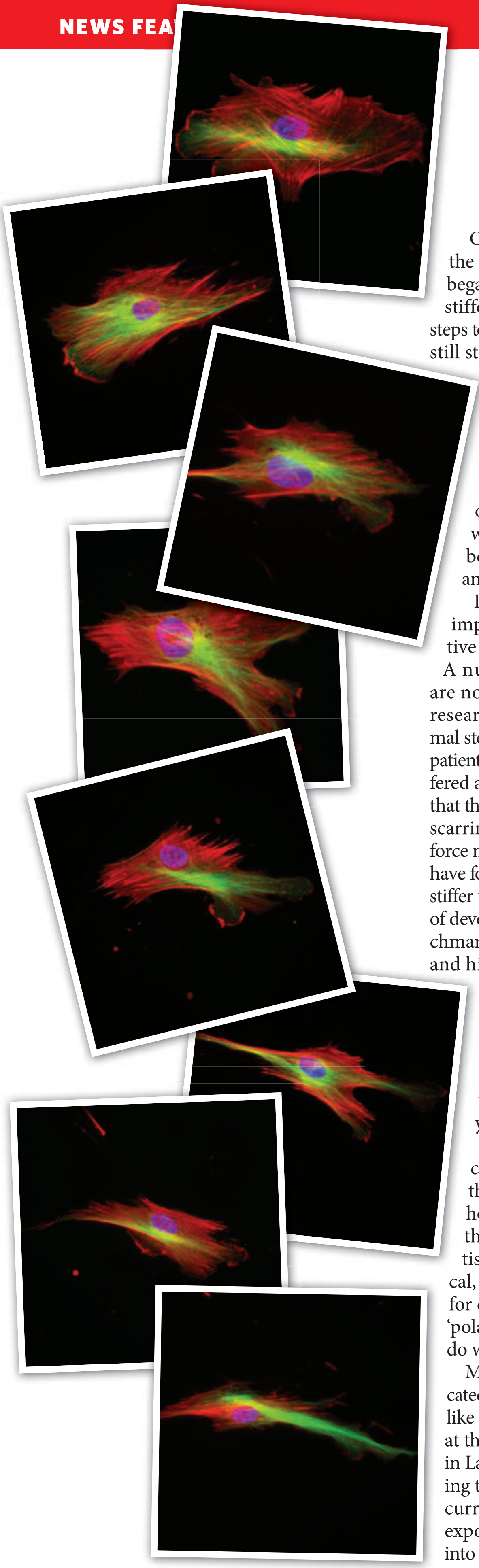
Eschewing growth factors

Monitoring the physical environment could also serve a purpose during development, by helping cells to detect where they are, and to migrate, divide or differentiate appropriately. This has required a particular change in thinking for many biologists. The field has long been dominated by the idea that developmental decision-making is directed by the chemical signals inherited from a cell's parent, or received from its neighbours and the environment. That way of thinking leads researchers to encourage stem cells to form heart muscle cells, neurons or other cell types simply by adding cocktails of proteins known as growth factors.

In 2006, cell and molecular biophysicist Dennis Discher at the University of Pennsylvania reported he had done away with growth factors and allowed force to trigger a change in cell fate[6]. His team studied mesenchymal stem cells, a kind of cell that normally grows in the soft, fatty bone marrow where it generates bone progenitors, but that is also thought to move to other locations and give rise to a range of cell types including nerve and muscle. They grew the cells on gels made from polyacrylamide and collagen that mimicked the softness of bone marrow. By changing the degree of chemical crosslinking in the polyacrylamide, the team was able to alter the stiffness of the gel to be more like that of different body tissues.

F. LEMMENS/ALAMY

Cells may share the quality of 'tensegrity' with the constructions of sculptor Kenneth Snelson.



Day in the life: a stiff substrate converts stem cells into bone cells over 24 hours.

On a relatively soft base that resembled the sponginess of brain tissue, the stem cells began to form the precursors of neurons; on stiffer, muscle-like substrates the cells took steps towards forming muscle stem cells; and on still stiffer substrates resembling developing bone they started to become bone-forming cells. The team still had to add growth factors to get the cells to differentiate fully, but by the time the stem cells had become comfortable for three weeks on their soft, medium or hard mattresses, growth factors that would normally get stem cells to switch between one developmental pathway and another had little effect.

Findings such as these have important implications for regenerative medicine, Discher says. A number of clinical trials are now under way in which researchers inject mesenchymal stem cells into the hearts of patients who have previously suffered a heart attack, in the hope that the cells will help repair the scarring caused by the attack. By using atomic force microscopy, Discher and his collaborators have found that in rats this kind of scar tissue is stiffer than normal heart muscle, resembling that of developing bone[7]. And last year, Bernd Fleischmann of the University of Bonn, Germany, and his colleagues found that mesenchymal stem cells formed bony spurs after they were injected into the scar tissue of damaged mouse hearts[8]. Although there are no published reports of human patients developing these problems, "You have to think about the microenvironment that you put cells in," Discher says.

> "The genome must be aware of the shape it is in charge of developing."
> — Emmanuel Farge

The new appreciation of cells' mechanical environment is also a complication in the lab. A number of studies have revealed how poorly cell-culture conditions mimic those that a cell would encounter in real tissue. Mammalian cells grown on a typical, flat, glass or plastic tissue culture dish, for example, do not develop the directional 'polarity' that they do in real tissue, but they do when grown in three-dimensional gels[9].

Many groups are now developing sophisticated cell-culture methods that have more life-like mechanics. Melody Swartz, a bioengineer at the Swiss Federal Institute of Technology in Lausanne (EPFL), and her team are studying the effects of 'interstitial flow': the gentle current to which almost all human cells are exposed as the fluid that bathes them seeps into the lymphatic system, the body's drainage network. The team has found that fibroblasts align themselves perpendicular to this flow during wound healing and inflammation. Swartz thinks that interstitial flow distributes chemical signals through tissue, and that cells may be so accustomed to its mechanical presence that they are lost without it. "They need the functional aspects — the mechanical environment — to act normally," she says. Her team has recently shown that the lymphatic cells themselves, which are notorious for losing many of their characteristic cell functions in culture, regain their character if she subjects them to flow.

"I don't think mechanobiology is a separate field from biology," says Swartz. "It's something you can either choose to consider or not, but it is always present." Not all labs are considering it — those that have switched to three-dimensional culture systems are still a minority, partly because cells grown that way are harder to study with standard microscopy and other techniques. But if mechanical stress has yet to be fully appreciated by cell biologists, it is becoming hard to ignore among developmental ones. As Ingber observed, nowhere are physical forces more apparent than in the developing embryo where tissues twist, fold and writhe into the beginnings of adult tissues and organs.

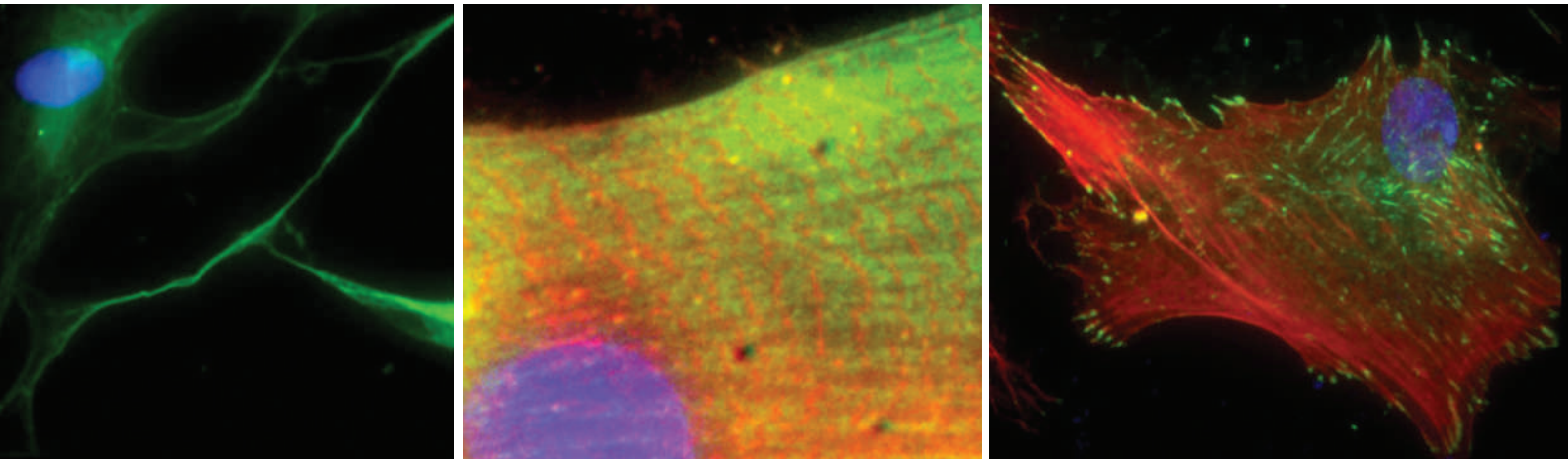
Let's twist again

In 2003, biophysicist and developmental biologist Emmanuel Farge at the Curie Institute in Paris gently squashed entire fruitfly embryos under a tiny sheet of glass, and found that they switched on a gene called *twist* in nearly all their cells, instead of the particular subsets where it should be expressed[10].

Farge has since shown that the normal compression of tissue that occurs as the embryo changes shape is required to switch on this gene in the nascent gut[11]. First he removed the cells that normally exert pressure on the gut as it grows, and showed that *twist* expression dropped. Then he devised a way to artificially exert a tiny force inside these embryos by injecting magnetic nanoparticles and tugging on them with an electromagnet. The expression of *twist* was restored.

Such studies challenge the idea that an embryo's shape and patterning is driven only by a genetic program, suggesting instead that shape and patterning can also drive gene expression. This mechanical control could provide cells with feedback on their changing position in the growing embryo. "The genome must be aware at key stages of the shape it is in charge of developing," says Farge. It also means that genes acting in two physically separate tissues — such as those that control tissue compression, and those that control *twist* — can still interact. This

A. ZAJAC



Mesenchymal stem cells grown on soft, medium or rigid matrices start developing into (left to right) neurons, muscle cells and bone cells.

A. ENGLER

kind of action at a distance, Farge speculates, might be involved in an embryo-wide, "global" system that coordinates development.

Given its importance in influencing cells in embryonic and adult tissues, it should come as no surprise that mechanical environments are now thought to contribute to disease. Even five years ago, says cell biologist Valerie Weaver, at the University of Pennsylvania, "there was a lot of scepticism" when she presented her work investigating links between tissue stiffness and cancer. "I know there was a little bit of snicker." But now a whole body of work has shown that manipulating the extracellular matrix in the cancer's microenvironment can switch normally dividing cells into excessively dividing, cancerous ones and vice versa[12].

Weaver and her colleagues investigated how mechanical forces on the outside of the cell could be converted into cancer-promoting signals on the inside — one of the key questions in the field[13]. Her team used a machine known as an electromechanical indentor, which presses on tissue in culture, to show that cancerous mammary tissue is stiffer than healthy tissue. Then they grew normal mammary cells in gels stiffened to different extents with collagen, so that they resembled either healthy tissue or tumours. Mammary cells grown in soft gels organized themselves into structures characteristic of normal breast tissue, whereas cells grown in stiff 'cancerous' gels did not. They found that the stiff gels pulled more on the membrane-spanning integrins, and this boosted the activity of an integrin-controlled signalling pathway that regulates tension in the cytoskeleton.

Crushing cancer

Weaver suggests that many oncogenes — genes which, when mutated, predispose a cell to becoming cancerous — can also activate biochemical pathways that increase internal tension. And in unpublished work, Weaver and her colleagues have shown that these changes may be some of the earliest events in cancer. They found that the extracellular matrix surrounding cells that harbour mutated oncogenes becomes stiffer even before those cells form invasive tumours, because the constitution of the extracellular matrix is changing. "That microenvironment is changing dramatically long before you get a tumour," says Weaver.

If her ideas are correct, then interfering with the mechanics of a cancer cell might override aberrations in its genes. And that's what Weaver found in her 2005 study: when she added chemicals that block integrin signalling to the cells grown in stiff gels, they grew into more normal-looking mammary-gland tissue. These signalling pathways might make a possible cancer drug target. And Ingber is pursuing the idea that cancer might be reversed by physically manipulating a tumour's microenvironment by, for example, implanting artificial materials that mimic the structural, mechanical and chemical properties of healthy tissue.

"Cells need the mechanical environment to act normally." — Melody Swartz

Ingber is also becoming more and more convinced that cells behave like Snelson's deformable sculptures. When he used a laser to break a single actin fibre, the shape of the entire cell altered to accommodate the change[3]. So, just like other tensegrity structures, the cytoskeleton can transmit a force from one point through the entire assembly. These mechanical signals are transmitted faster than a chemical signal can diffuse across a cell[14], and Ingber thinks that such mechanical changes may physically deform other molecules attached to the cytoskeleton, including many associated with the cell's metabolism. This cascade of events could eventually cause a dramatic change in cell behaviour[15].

Ingber goes even further, suggesting that living creatures are a Russian doll of tensegrities. On the largest scale, muscles tense against bones. Inside the body, the forces in these and other tissues are picked up by integrins and relayed to the cytoskeleton. The cytoskeleton itself is made up of macromolecular structures that are themselves tensegrities at the molecular level. Such a hierarchy of systems, he says, could explain how mechanical signals such as gravity or movement are transmitted from the macro to the micro and nano scales.

In the past two years, Ingber may have found the means to test his ideas. In 2007 he received seed money to establish a Harvard Institute for Biologically Inspired Engineering. And in October 2008, philanthropist Hansjörg Wyss gave US$125 million towards it. Looking back, Ingber says, "I was lucky to have my first 'Aha moment' as an undergraduate at a liberal arts college. In the United States, undergraduates can explore any field, and hence know no bounds between biology, chemistry, physics, architecture or art."

It is a boundlessness he tries to convey to the students he teaches today. Those who visit his office find that it is filled with tensegrity models. He even has a small Snelson sculpture on the window sill, one that the artist gave him. And in the classroom, some Harvard undergraduates are already exploring how to develop expandable lightweight structures based on tensegrity and cell architecture. But these are not designs for beautiful sculptures — they are water carriers for the developing world. ■

Claire Ainsworth is a freelance writer based in Southampton, UK.

1. His, W. *R. Soc. Edinburgh Proc.* **15,** 287–298 (1888).
2. Folkman, J. & Moscona, A. *Nature* **273,** 345–349 (1978).
3. Kumar, S. *et al. Biophys. J.* **90,** 3762–3773 (2006).
4. Solon, J., Levental, I., Sengupta, K., Georges, P. C. & Janmey, P. A. *Biophys. J.* **93,** 4453–4461 (2007).
5. Saxena, V. *et al. Plast. Reconstr. Surg.* **114,** 1086–1096 (2004).
6. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. *Cell* **126,** 677–689 (2006).
7. Berry, M. F. *et al. Am. J. Physiol. Heart Circ. Physiol.* **290,** H2196–H2203 (2006).
8. Breitbach, M. *et al. Blood* **110,** 1362–1369 (2007).
9. Pampaloni, F., Reynaud, E. G. & Stelzer, E. H. *Nature Rev. Mol. Cell Biol.* **8,** 839–845 (2007).
10. Farge E. *Curr. Biol.* **13,** 1365–1377 (2003).
11. Desprat, N., Supatto, W., Pouille, P. A., Beaurepaire, E. & Farge, E. *Dev. Cell* **15,** 470–477 (2008).
12. Albini, A. & Sporn, M. B. *Nature Rev. Cancer* **7,** 139–147 (2007).
13. Paszek, M. J. *et al. Cancer Cell* **8,** 241–254 (2005).
14. Na, S. *et al. Proc. Natl Acad. Sci. USA* **105,** 6626–6631 (2008).
15. Ingber, D. E. *Prog. Biophys. Mol. Biol.* **97,** 163–179 (2008).

CORRESPONDENCE

Bovine TB: don't get rid of the cat because the mice have gone

SIR — Paul Torgerson and David Torgerson argue in Correspondence that research and surveillance for tuberculosis (TB) in cattle can only be justified on animal-health grounds because bovine TB poses a negligible threat to human health ('Does risk to humans justify high cost of fighting bovine TB?' *Nature* **455,** 1029; 2008). But the relationship between surveillance of cattle and the risk of humans contracting the disease is not as simple as they imply.

The current surveillance to remove infected cattle at an early stage in the disease process reduces the prevalence and, more important, curtails the course of the disease in individual cattle. This is particularly relevant in areas of higher incidence, where testing is an annual event. For example, three UK counties breached 10% herd incidence of confirmed cases in 2007.

Without a test-and-cull policy and abattoir surveillance, bovine TB would advance to its more infectious stages and lead to an increased risk of transmission of this primarily airborne disease to humans. Removing animals at an early stage and controlling the prevalence of the disease in the national herd is therefore likely to reduce the probability of transmission to humans.

We would argue that active surveillance for bovine TB followed by the removal of infected cattle is part of the reason why this disease is not seen more often in humans. When using a cat to control mice, you shouldn't get rid of the cat just because you don't see any more mice.

We agree that protecting animal health and reducing losses in animal productivity are important justifications for surveillance and research of bovine TB. But keeping the risk of human exposure as low as possible must also be factored into the equation.

Noel H. Smith Veterinary Laboratories Agency, Weybridge, Addlestone, Surrey KT15 3NB, UK, and Centre for the Study of Evolution, John Maynard Smith Building,University of Sussex, Brighton BN1 9QG, UK
e-mail: n.h.smith@vla.defra.gsi.gov.uk
Richard Clifton-Hadley Veterinary Laboratories Agency, Weybridge, Addlestone, Surrey, KT15 3NB, UK

Bovine TB: stopping disease control would block all live exports

SIR — In their Correspondence, Paul Torgerson and David Torgerson question the importance of controlling bovine tuberculosis (TB) in the United Kingdom ('Does risk to humans justify high cost of fighting bovine TB?' *Nature* **455,** 1029; 2008). However, they disregard the crucial role of disease control in protecting international trade — a key reason why the UK government is trying to eradicate bovine TB.

To be exported from the United Kingdom, cattle aged 42 days or older must have tested negative for bovine TB within the previous 30 days. In 1995, the value of live cattle exports was £77.6 million (US$115 million); after the imposition of the ban on exports because of BSE (bovine spongiform encephalopathy), this fell to zero (see http://tinyurl.com/6j8o2w). Although the ban on exports from the United Kingdom was lifted in 2006, live cattle exports are only just now starting to recover.

To stop control of bovine TB would once again block all live exports. UK government expenditure on the foot-and-mouth disease outbreak in 2001 was £2.79 billion, and the protection of international trade was a key driver of the control measures used at that time (see http://tinyurl.com/67p9g5).

The protection of trade status is one aspect of an integrated policy for controlling bovine TB that safeguards public health and the agricultural economy as well as animal health and welfare. Indeed, as we face an increasing problem with bovine TB in the United Kingdom, the success of this control strategy in protecting public health is evidenced by the sporadic nature of *Mycobacterium bovis* infections in the human population (see A. L. Gibson *et al. J. Clin Microbiol.* **42,** 431–434; 2004; J. T. Evans *et al. Lancet* **369,** 1270–1276; 2007). Now is not the time to dismantle controls for bovine TB.

Stephen V. Gordon UCD College of Life Sciences and UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, Ireland
e-mail: stephen.gordon@ucd.ie

Right environment can enhance 'innate' entrepreneurial skills

SIR — In their Commentary 'The innovative brain' (*Nature* **456,** 168–169), Barbara Sahakian and her colleagues raise the important question of whether entrepreneurial skills can be taught. In my experience, an entrepreneur's 'cold' skills (in creating a compelling business plan, for example) are easier to teach than the more intuitive 'hot' decision-making abilities that stimulate success.

Successful entrepreneurs have enormous energy, think independently and show inspiring leadership qualities that help to realize a shared vision. However, their flair for weighing up and pursuing the best opportunities is honed by past experience — for example, in spotting promising factors in new companies. An entrepreneur's natural ability is founded in the interaction of genes and environment.

During the past 20 years I have participated in the development of the entrepreneurial environment here in Cambridge. In 1997, only 17% of the entrepreneurs in our portfolio companies were 'serial entrepreneurs' — people who already had a record of achievement. Now this percentage has risen to 70%, which reflects the rapid increase in entrepreneurial talent in Britain. Today there are more than 1,000 high-tech companies in Cambridge alone, with at least one entrepreneur involved in each of them.

Know-how is transmitted 'in the air' within these high-technology clusters. They are also valuable in tolerating risk and possible failure well: as bioscience entrepreneur Andy Richards has pointed out, Cambridge has become a low-risk area for doing high-risk things. This entrepreneurial environment promotes venturesome behaviour in the face of risky business opportunity. Failure carries no stigma when it is recognized as a necessary part of the entrepreneurial process — as it is here — given that some 30% of all new companies will fail anyway.

Entrepreneurial activities and innovative ideas are particularly needed in today's economic climate, both for the wealth of the United Kingdom and for the global economy. Understanding how the innate components of these abilities can be enhanced will help to educate the next generation of entrepreneurs for success.

Hermann Hauser, Amadeus Capital, Mount Pleasant House, 2 Mount Pleasant, Cambridge CB3 0RN, UK
e-mail: hhauser@amadeuscapital.com

Readers are welcome to comment at http://tinyurl.com/5opr6g

Contributions may be submitted to correspondence@nature.com. Correspondence should be signed by no more than three authors; preferably by one. Published contributions are edited. Science publishing issues are regularly featured at Nautilus (http://blogs.nature.com/nautilus), where we welcome comments and debate.

nature

COMMENTARY

The food crisis isn't over

Although the credit crunch has lowered the price of food, a global recession now raises the hunger pains of the most vulnerable. The stage is set for the next international food crisis, says **Joachim von Braun**.

Just a few months ago, the 'crisis' on everyone's mind was the spike in food prices. Then the financial crisis gripped the world, food prices declined somewhat, and agricultural concerns lost their sense of immediacy. But we should not be complacent. The credit crunch is likely to have strong and long-lasting effects on emerging economies and on the people most in need.

RECESSION WATCH

Between 2005 and summer 2008, the international prices of wheat and maize (corn) tripled, and that of rice grew fivefold. Poor people typically spend 50–70% of their income on food, and their wages did not adjust quickly enough to compensate for their shrinking purchasing power. From 2003–05 to 2007, the number of undernourished people increased from some 848 million to 923 million, largely because of the food-price crisis[1].

The worldwide credit crunch has let some air out of the commodity price bubble, providing a little relief — cereal prices have fallen by about 30–40% in recent months. But recession also threatens to cut the income and employment of the most vulnerable and undermine investment in agricultural production. The economic bailout and suggested market regulations now being discussed will not protect food prices from future spikes. The world's food worries are by no means over.

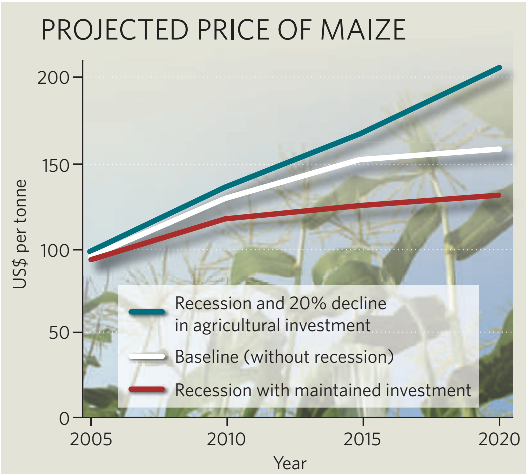
The path to malnourishment

An International Food Policy Research Institute (IFPRI) model, developed by division director Mark Rosegrant, explores what might happen in the face of the recession. If global economic annual growth falls by 2–3 percentage points below recent years' figure of about 5%, and agricultural investment declines in parallel by 20% — a realistic scenario — this would result in cereal prices 30% above what is expected without a recession by 2020 (ref. 2). Globally, 16 million more children would be malnourished.

However, if spending on agricultural research and development (R&D) is maintained (assuming a modest 3% decline in investment growth) in the face of this recession, cereal prices would be about 15% lower than the non-recession baseline in 2020, and malnutrition would be about the same as in the baseline scenario (see graph).

Agricultural R&D is one of the most effective types of investment for preventing food crises, promoting economic growth and reducing poverty[3]. Global underinvestment has impaired productivity in recent years. As growth of public agricultural R&D spending in low- and middle-income countries decreased from 3% in the 1980s to 1.9% in the 1990s (ref. 4), global grain productivity growth declined from about 3% to 1% (ref. 5).

Doubling all agricultural R&D in developing countries between 2008 and 2013, from US$5 billion to $10 billion, could increase agricultural growth by 1.1 percentage points a year, and lift about 282 million people out of poverty by 2020 (ref. 2). Although this would mark a historic turnaround in such investment, it pales in comparison to the financial bailout costs.



Much investment will have to be facilitated by public sources, in view of the deepening shortage of private capital and the risky investment climate. Institutions such as the World Bank and the Asian Development Bank have recently reversed their declining investments in agriculture, with substantial increases seen in 2007 and 2008. Such trends need to continue.

Investment in R&D is necessary to keep long-term food prices at reasonable levels. But to avoid price bubbles, action needs to be taken to improve market efficiency and food trade in crisis situations. The IFPRI recommends an emergency reserve of 300,000 tonnes of grain, located at strategic points in the developing world, that would be overseen by the United Nations World Food Programme. Current emergency grain reserves are controlled nationally and lack international coordination.

In addition, a global intelligence unit should be created to advise when market intervention is needed. Should grain prices be driven into a price bubble again, an international body could act through the futures market — effectively promising to sell grain at low prices in the future. In most cases, this should cause prices to decline before sells are realized, so no actual money need change hands. In the unusual circumstance in which this futures activity would need to be paid for, the international body would have to draw on a 'virtual reserve' — a promise of funds from participating nations[6]. After discussion at the 2008 G8 (Group of Eight) summit of the world's leading industralized nations, such a scheme is now under consideration at high political levels.

At the same time, subsidies for biofuels, which have diverted funds from food and feed production, should be removed. Although biofuels will surely play a part in future energy systems, the billions spent on subsidies would be better invested in R&D for innovative biomass use and other energy innovations. Also required is a scaling up of social programmes, including employment programmes, school feeding, and early childhood nutrition interventions.

The successful resolution of the food crisis should not be measured by the drop of commodity prices to previous levels, but by sufficient food on poor people's plates today and long into the future. ■

Joachim von Braun is the director general of the International Food Policy Research Institute in Washington DC.
e-mail: j.vonbraun@cgiar.org

1. *Hunger on the Rise: Soaring Prices Add 75 Million People to Global Hunger Rolls.* Briefing paper (FAO, 2008).
2. von Braun, J. *Food and Financial Crises: Implications for Agriculture and the Poor* (IFPRI, 2008).
3. Fan, S., Gulati, A. & Thorat, S. *Agric. Econ.* **39,** 163–170 (2008).
4. Beintema, N. M. & Stads, G.-J. *Measuring Agricultural Research Investments: A Revised Global Picture* (IFPRI, 2008).
5. *World Development Report 2008: Agriculture for Development* (World Bank, 2007).
6. von Braun, J. & Torero, M. *Physical and Virtual Global Food Reserves to Protect the Poor and Prevent Market Failure.* IFPRI Policy Brief 4 (IFPRI, 2008).

COMMENTARY

Towards responsible use of cognitive-enhancing drugs by the healthy

Society must respond to the growing demand for cognitive enhancement. That response must start by rejecting the idea that 'enhancement' is a dirty word, argue **Henry Greely and colleagues**.

Today, on university campuses around the world, students are striking deals to buy and sell prescription drugs such as Adderall and Ritalin — not to get high, but to get higher grades, to provide an edge over their fellow students or to increase in some measurable way their capacity for learning. These transactions are crimes in the United States, punishable by prison.

Many people see such penalties as appropriate, and consider the use of such drugs to be cheating, unnatural or dangerous. Yet one survey[1] estimated that almost 7% of students in US universities have used prescription stimulants in this way, and that on some campuses, up to 25% of students had used them in the past year. These students are early adopters of a trend that is likely to grow, and indications suggest that they're not alone[2].

In this article, we propose actions that will help society accept the benefits of enhancement, given appropriate research and evolved regulation. Prescription drugs are regulated as such not for their enhancing properties but primarily for considerations of safety and potential abuse. Still, cognitive enhancement has much to offer individuals and society, and a proper societal response will involve making enhancements available while managing their risks.

C. GALLAGHER/SPL

Adderall is one of several drugs increasingly used to enhance cognitive function.

Paths to enhancement

Many of the medications used to treat psychiatric and neurological conditions also improve the performance of the healthy. The drugs most commonly used for cognitive enhancement at present are stimulants, namely Ritalin (methylphenidate) and Adderall (mixed amphetamine salts), and are prescribed mainly for the treatment of attention deficit hyperactivity disorder (ADHD). Because of their effects on the catecholamine system, these drugs increase executive functions in patients and most healthy normal people, improving their abilities to focus their attention, manipulate information in working memory and flexibly control their responses[3]. These drugs are widely used therapeutically. With rates of ADHD in the range of 4–7% among US college students using DSM criteria[4], and stimulant medication the standard therapy, there are plenty of these drugs on campus to divert to enhancement use.

A newer drug, modafinil (Provigil), has also shown enhancement potential. Modafinil is approved for the treatment of fatigue caused by narcolepsy, sleep apnoea and shift-work sleep disorder. It is currently prescribed off label for a wide range of neuropsychiatric and other medical conditions involving fatigue[5] as well as for healthy people who need to stay alert and awake when sleep deprived, such as physicians on night call[6]. In addition, laboratory studies have shown that modafinil enhances aspects of executive function in rested healthy adults, particularly inhibitory control[7]. Unlike Adderall and Ritalin, however, modafinil prescriptions are not common, and the drug is consequently rare on the college black market. But anecdotal evidence and a readers' survey both suggest that adults sometimes obtain modafinil from their physicians or online for enhancement purposes[2].

A modest degree of memory enhancement is possible with the ADHD medications just mentioned as well as with medications developed for the treatment of Alzheimer's disease such as Aricept (donepezil), which raise levels of acetylcholine in the brain[8]. Several other compounds with different pharmacological actions are in early clinical trials, having shown positive effects on memory in healthy research subjects (see, for example, ref. 9). It is too early to know whether any of these new drugs will be proven safe and effective, but if one is it will surely be sought by healthy middle-aged and elderly people contending with normal age-related memory decline, as well as by people of all ages preparing for academic or licensure examinations.

Favouring innovation

Human ingenuity has given us means of enhancing our brains through inventions such as written language, printing and the Internet. Most authors of this Commentary are teachers and strive to enhance the minds of their students, both by adding substantive information and by showing them new and better ways to process that information. And we are all aware of the abilities to enhance our brains with adequate exercise, nutrition and sleep. The drugs just reviewed, along with newer technologies such as brain stimulation and prosthetic brain chips, should be viewed in the same general category as education, good health habits, and information technology — ways that our uniquely innovative species tries to improve itself.

Of course, no two enhancements are equivalent in every way, and some of the differences have moral relevance. For example, the benefits of education require some effort at self-improvement whereas the benefits of sleep do not. Enhancing by nutrition involves changing what we ingest and is therefore invasive in a way

that reading is not. The opportunity to benefit from Internet access is less equitably distributed than the opportunity to benefit from exercise. Cognitive-enhancing drugs require relatively little effort, are invasive and for the time being are not equitably distributed, but none of these provides reasonable grounds for prohibition. Drugs may seem distinctive among enhancements in that they bring about their effects by altering brain function, but in reality so does any intervention that enhances cognition. Recent research has identified beneficial neural changes engendered by exercise[10], nutrition[11] and sleep[12], as well as instruction[13] and reading[14]. In short, cognitive-enhancing drugs seem morally equivalent to other, more familiar, enhancements.

Many people have doubts about the moral status of enhancement drugs for reasons ranging from the pragmatic to the philosophical, including concerns about short-circuiting personal agency and undermining the value of human effort[15]. Kass[16], for example, has written of the subtle but, in his view, important differences between human enhancement through biotechnology and through more traditional means. Such arguments have been persuasively rejected (for example, ref. 17). Three arguments against the use of cognitive enhancement by the healthy quickly bubble to the surface in most discussions: that it is cheating, that it is unnatural and that it amounts to drug abuse.

In the context of sports, pharmacological performance enhancement is indeed cheating. But, of course, it is cheating because it is against the rules. Any good set of rules would need to distinguish today's allowed cognitive enhancements, from private tutors to double espressos, from the newer methods, if they are to be banned.

> "We should welcome new methods of improving our brain function."

As for an appeal to the 'natural', the lives of almost all living humans are deeply unnatural; our homes, our clothes and our food — to say nothing of the medical care we enjoy — bear little relation to our species' 'natural' state. Given the many cognitive-enhancing tools we accept already, from writing to laptop computers, why draw the line here and say, thus far but no further?

As for enhancers' status as drugs, drug abuse is a major social ill, and both medicinal and recreational drugs are regulated because of possible harms to the individual and society. But drugs are regulated on a scale that subjectively judges the potential for harm from the very dangerous (heroin) to the relatively harmless (caffeine). Given such regulation, the mere fact that cognitive enhancers are drugs is no reason to outlaw them.

Based on our considerations, we call for a presumption that mentally competent adults should be able to engage in cognitive enhancement using drugs.

Substantive concerns and policy goals

All technologies have risks as well as benefits. Although we reject the arguments against enhancement just reviewed, we recognize at least three substantive ethical concerns.

The first concern is safety. Cognitive enhancements affect the most complex and important human organ, and the risk of unintended side effects is therefore both high and consequential. Although regulations governing medicinal drugs ensure that they are safe and effective for their therapeutic indications, there is no equivalent vetting for unregulated 'off label' uses, including enhancement uses. Furthermore, acceptable safety in this context depends on the potential benefit. For example, a drug that restored good cognitive functioning to people with severe dementia but caused serious adverse medical events might be deemed safe enough to prescribe, but these risks would be unacceptable for healthy individuals seeking enhancement.

Enhancement in children raises additional issues related to the long-term effects on the developing brain. Moreover, the possibility of raising cognitive abilities beyond their species-typical upper bound may engender new classes of side effects. Persistence of unwanted recollections, for example, has dearly negative effects on the psyche[18].

An evidence-based approach is required to evaluate the risks and benefits of cognitive enhancement. At a minimum, an adequate policy should include mechanisms for the assessment of both risks and benefits for enhancement uses of drugs and devices, with special attention to long-term effects on development and to the possibility of new types of side effects unique to enhancement. But such considerations should not lead to an insistence on higher thresholds than those applied to medications.

We call for an evidence-based approach to the evaluation of the risks and benefits of cognitive enhancement.

The second concern is freedom, specifically freedom from coercion to enhance. Forcible medication is generally reserved for rare cases in which individuals are deemed threats to themselves or others. In contrast, cognitive enhancement in the form of education is required for almost all children at some substantial cost to their liberty, and employers are generally free to require employees to have certain educational credentials or to obtain them. Should schools and employers be allowed to require pharmaceutical enhancement as well? And if we answer 'no' to this question, could coercion occur indirectly, by the need to compete with enhanced classmates and colleagues?

Questions of coercion and autonomy are particularly acute for military personnel and for children. Soldiers in the United States and elsewhere have long been offered stimulant medications including amphetamine and modafinil to enhance alertness, and in the United States are legally required to take medications if ordered to for the sake of their military performance[19]. For similar reasons, namely the safety of the individual in question and others who depend on that individual in dangerous situations, one could imagine other occupations for which enhancement might be justifiably required. A hypothetical example is an extremely safe drug that enabled surgeons to save more patients. Would it be wrong to require this drug for risky operations?

Appropriate policy should prohibit coercion except in specific circumstances for specific occupations, justified by substantial gains in safety. It should also discourage indirect coercion. Employers, schools or governments should not generally require the use of cognitive enhancements. If particular enhancements are shown to be sufficiently safe and effective, this position might be revisited for those interventions.

Children once again represent a special case as they cannot make their own decisions. Comparisons between estimates of ADHD prevalence and prescription numbers have led some to suspect that children in certain school districts are taking enhancing drugs at the behest of achievement-oriented parents, or teachers seeking more orderly classrooms[20]. Governments may be willing to let competent adults take certain risks for the sake of enhancement while restricting the ability to take such risky decisions on behalf of children.

The third concern is fairness. Consider an examination that only a certain percentage can pass. It would seem unfair to allow some, but not all, students to use cognitive enhancements, akin to allowing some students taking a maths test to use a calculator while others must go without. (Mitigating such unfairness may raise issues of indirect coercion, as discussed above.) Of course, in some ways, this kind of unfairness already exists. Differences in education, including private tutoring, preparatory courses and other enriching experiences give some students an advantage over others.

Whether the cognitive enhancement is substantially unfair may depend on its availability, and on the nature of its effects. Does it actually improve learning or does it just temporarily boost exam performance? In the latter

case it would prevent a valid measure of the competency of the examinee and would therefore be unfair. But if it were to enhance long-term learning, we may be more willing to accept enhancement. After all, unlike athletic competitions, in many cases cognitive enhancements are not zero-sum games. Cognitive enhancement, unlike enhancement for sports competitions, could lead to substantive improvements in the world.

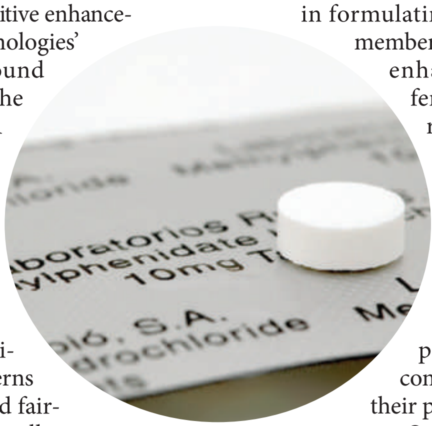
Fairness in cognitive enhancements has a dimension beyond the individual. If cognitive enhancements are costly, they may become the province of the rich, adding to the educational advantages they already enjoy. One could mitigate this inequity by giving every exam-taker free access to cognitive enhancements, as some schools provide computers during exam week to all students. This would help level the playing field.

Policy governing the use of cognitive enhancement in competitive situations should avoid exacerbating socioeconomic inequalities, and should take into account the validity of enhanced test performance. In developing policy for this purpose, problems of enforcement must also be considered. In spite of stringent regulation, athletes continue to use, and be caught using, banned performance-enhancing drugs.

We call for enforceable policies concerning the use of cognitive-enhancing drugs to support fairness, protect individuals from coercion and minimize enhancement-related socioeconomic disparities.

Maximum benefit, minimum harm

The new methods of cognitive enhancement are 'disruptive technologies' that could have a profound effect on human life in the twenty-first century. A laissez-faire approach to these methods will leave us at the mercy of powerful market forces that are bound to be unleashed by the promise of increased productivity and competitive advantage. The concerns about safety, freedom and fairness, just reviewed, may well seem less important than the attractions of enhancement, for sellers and users alike.



The prescription drug Ritalin is illegally traded among students.

Motivated by some of the same considerations, Fukuyama[21] has proposed the formation of new laws and regulatory structures to protect against the harms of unrestrained biotechnological enhancement. In contrast, we suggest a policy that is neither laissez-faire nor primarily legislative. We propose to use a variety of scientific, professional, educational and social resources, in addition to legislation, to shape a rational, evidence-based policy informed by a wide array of relevant experts and stakeholders. Specifically, we propose four types of policy mechanism.

The first mechanism is an accelerated programme of research to build a knowledge base concerning the usage, benefits and associated risks of cognitive enhancements. Good policy is based on good information, and there is currently much we do not know about the short- and long-term benefits and risks of the cognitive-enhancement drugs currently being used, and about who is using them and why. For example, what are the patterns of use outside of the United States and outside of college communities? What are the risks of dependence when used for cognitive enhancement? What special risks arise with the enhancement of children's cognition? How big are the effects of currently available enhancers? Do they change 'cognitive style', as well as increasing how quickly and accurately we think? And given that most research so far has focused on simple laboratory tasks, how do they affect cognition in the real world? Do they increase the total knowledge and understanding that students take with them from a course? How do they affect various aspects of occupational performance?

We call for a programme of research into the use and impacts of cognitive-enhancing drugs by healthy individuals.

The second mechanism is the participation of relevant professional organizations in formulating guidelines for their members in relation to cognitive enhancement. Many different professions have a role in dispensing, using or working with people who use cognitive enhancers. By creating policy at the level of professional societies, it will be informed by the expertise of these professionals, and their commitment to the goals of their profession.

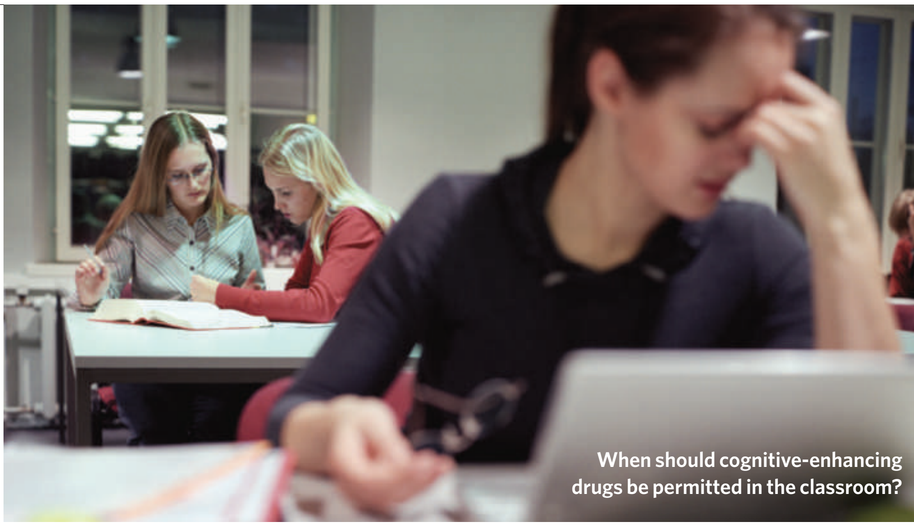
One group to which this recommendation applies is physicians, particularly in primary care, paediatrics and psychiatry, who are most likely to be asked for cognitive enhancers. These physicians are sometimes asked to prescribe for enhancement by patients who exaggerate or fabricate symptoms of ADHD, but they also receive frank requests, as when a patient says "I know I don't meet diagnostic criteria for ADHD, but I sometimes have trouble concentrating and staying organized, and it would help me to have some Ritalin on hand for days when I really need to be on top of things at work." Physicians who view medicine as devoted to healing will view such prescribing as inappropriate, whereas those who view medicine more broadly as helping patients live better or achieve their goals would be open to considering such a request[22]. There is certainly a precedent for this broader view in certain branches of medicine, including plastic surgery, dermatology, sports medicine and fertility medicine.

Because physicians are the gatekeepers to medications discussed here, society looks to them for guidance on the use of these medications and devices, and guidelines from other professional groups will need to take into account the gatekeepers' policies. For this reason, the responsibilities that physicians bear for the consequences of their decisions are particularly sensitive, being effectively decisions for all of us. It would therefore be helpful if physicians as a profession gave serious consideration to the ethics of appropriate prescribing of cognitive enhancers, and consulted widely as to how to strike the balance of limits for patient benefit and protection in a liberal democracy. Examples of such limits in other areas of enhancement medicine include the psychological screening of candidates for cosmetic surgery or tubal ligation, and upper bounds on maternal age or number of embryos transferred in fertility treatments. These examples of limits may not be specified by law, but rather by professional standards.

Other professional groups to which this recommendation applies include educators and human-resource professionals. In different ways, each of these professions has responsibility for fostering and evaluating cognitive performance and for advising individuals who are seeking to improve their performance, and some responsibility also for protecting the interests of those in their charge. In contrast to physicians, these professionals have direct conflicts of interest that must be addressed in whatever guidelines they recommend: liberal use of cognitive enhancers would be expected to encourage classroom order and raise standardized measures of student achievement, both of which are in the interests of schools; it would also be expected to promote workplace productivity, which is in the interests of employers.

Educators, academic admissions officers and credentials evaluators are normally responsible for ensuring the validity and integrity of their examinations, and should be tasked with formulating policies concerning enhancement by test-takers. Laws pertaining to testing accommodations for people with disabilities provide

SPL


MIKA/ZEFA/CORBIS

When should cognitive-enhancing drugs be permitted in the classroom?

a starting point for discussion of some of the key issues, such as how and when enhancements undermine the validity of a test result and the conditions under which enhancement should be disclosed by a test-taker.

The labour and professional organizations of individuals who are candidates for on-the-job cognitive enhancement make up our final category of organization that should formulate enhancement policy. From assembly line workers to surgeons, many different kinds of employee may benefit from enhancement and want access to it, yet they may also need protection from the pressure to enhance.

We call for physicians, educators, regulators and others to collaborate in developing policies that address the use of cognitive-enhancing drugs by healthy individuals.

The third mechanism is education to increase public understanding of cognitive enhancement. This would be provided by physicians, teachers, college health centres and employers, similar to the way that information about nutrition, recreational drugs and other public-health information is now disseminated. Ideally it would also involve discussions of different ways of enhancing cognition, including through adequate sleep, exercise and education, and an examination of the social values and pressures that make cognitive enhancement so attractive and even, seemingly, necessary.

We call for information to be broadly disseminated concerning the risks, benefits and alternatives to pharmaceutical cognitive enhancement.

The fourth mechanism is legislative. Fundamentally new laws or regulatory agencies are not needed. Instead, existing law should be brought into line with emerging social norms and information about safety. Drug law is one of the most controversial areas of law, and it would be naive to expect rapid or revolutionary change in the laws governing the use of controlled substances. Nevertheless, these laws should be adjusted to avoid making felons out of those who seek to use safe cognitive enhancements. And regulatory agencies should allow pharmaceutical companies to market cognitive-enhancing drugs to healthy adults provided they have supplied the necessary regulatory data for safety and efficacy.

We call for careful and limited legislative action to channel cognitive-enhancement technologies into useful paths.

Conclusion

Like all new technologies, cognitive enhancement can be used well or poorly. We should welcome new methods of improving our brain function. In a world in which human workspans and lifespans are increasing, cognitive enhancement tools — including the pharmacological — will be increasingly useful for improved quality of life and extended work productivity, as well as to stave off normal and pathological age-related cognitive declines[23]. Safe and effective cognitive enhancers will benefit both the individual and society.

"Many kinds of employee may benefit from enhancement."

But it would also be foolish to ignore problems that such use of drugs could create or exacerbate. With this, as with other technologies, we need to think and work hard to maximize its benefits and minimize its harms. ■

Henry Greely is at Stanford Law School, Crown Quadrangle, 559 Nathan Abbott Way, Stanford, California 94305-8610, USA.
e-mail: hgreely@stanford.edu
Barbara Sahakian is at the Department of Psychiatry, University of Cambridge, and MRC/Wellcome Trust Behavioural and Clinical Neuroscience Institute, Cambridge, UK.
e-mail: jenny.hall@cpft.nhs.uk
John Harris is at the Institute for Science, Ethics and Innovation, and Wellcome Strategic Programme in The Human Body, its Scope, Limits and Future, University of Manchester, Oxford Road, Manchester M13 9PL, UK.
e-mail: john.harris@manchester.ac.uk
Ronald C. Kessler is at Harvard Medical School, Department of Health Care Policy, 180 Longwood Avenue, Boston, Massachusetts 02115-5899, USA.
e-mail: kessler@hcp.med.harvard.edu
Michael Gazzaniga is at the Sage Center for the Study of Mind, University of California, Santa Barbara, California 93106-9660, USA.
e-mail: m.gazzaniga@psych.ucsb.edu
Philip Campbell is at *Nature*, 4 Crinan St, London N1 9XW, UK.
e-mail: nature@nature.com
Martha J. Farah is at the Center for Cognitive Neuroscience, University of Pennsylvania, 3720 Walnut Street, Room B51, Philadelphia, Pennsylvania 19104-6241, USA.
e-mail: mfarah@psych.upenn.edu

1. McCabe, S. E., Knight, J. R., Teter, C. J. & Wechsler, H. *Addiction* **100,** 96–106 (2005).
2. Maher, B. *Nature* **452,** 674–675 (2008).
3. Sahakian, B. & Morein-Zamir, S. *Nature* **450,** 1157–1159 (2007).
4. Weyandt, L. L & DuPaul, G. *J. Atten. Disord.* **10,** 9–19 (2006).
5. Minzenberg, M. J. & Carter, C. S. *Neuropsychopharmacology* **33,** 1477–1502 (2008).
6. Vastag, B. *J. Am. Med. Assoc.* **291,** 167–170 (2004).
7. Turner, D. C. *et al. Psychopharmacology* **165,** 260–269 (2003).
8. Grön, G., Kirstein, M., Thielscher, A., Riepe, M. W. & Spitzer, M. *Psychopharmacology* **182,** 170–179 (2005).
9. Lynch, G. & Gall C. M. *Trends Neurosci.* **29,** 554–562 (2006).
10. Hillman, C. H., Erikson, K. I. & Kramer, A. F. *Nature Rev. Neurosci.* **9,** 58-65 (2008).
11. Almeida, S. S. *et al. Nutr. Neurosci.* **5,** 311–320 (2002).
12. Boonstra, T. W., Stins, J. F., Daffertshofer, A. & Beek, P. J. *Cell. Mol. Life Sci.* **64,** 934–946 (2007).
13. Draganski, B. *et al. Nature* **427,** 311–312 (2004).
14. Schlaggar, B. L. & McCandliss, B. D. *Annu. Rev. Neurosci.* **30,** 475–503 (2007).
15. Farah, M. J. *et al. Nature Rev. Neurosci.* **5,** 421–425 (2004).
16. Kass, L. R. *et al. Beyond Therapy: Biotechnology and the Pursuit of Happiness* (President's Council on Bioethics, 2003); available at www.bioethics.gov/reports/beyondtherapy.
17. Harris, J. *Enhancing Evolution: The Ethical Case for Making Better People* (Princeton Univ. Press, 2007).
18. Schacter, D. L. *Seven Sins of Memory* (Houghton Mifflin, 2002).
19. Moreno, J. D. *Mind Wars: Brain Research and National Defense* (Dana Press, 2006).
20. Diller, L. H. *Hastings Cent. Rep.* **26,** 12–18 (1996).
21. Fukuyama, F. *Our Posthuman Future: Consequences of the Biotechnology Revolution* (Farrar, Straus and Giroux, 2002).
22. Chatterjee, A. *Neurology* **63,** 968–974 (2004).
23. Beddington, J. *et al. Nature* **455,** 1057–1060 (2008).

Acknowledgement This article is the result of a seminar held by the authors at Rockefeller University. Funds for the seminar were provided by Rockefeller University and *Nature*.

Competing interests B.S. consults for a number of pharmaceutical companies and Cambridge Cognition, and holds shares in CeNeS. R.C.K. consults for and has received grants from a number of pharmaceutical companies.

Join the debate on this topic at Nature Network
➧ http://tinyurl.com/6nyu29

improving a standardized technical core that already exists.

Scientific understanding related to a standardized core is much easier to apply than science aimed at elucidating the theoretical foundations, causes or dynamics of a problem. When knowledge is not largely embodied in an effective technology, but must instead be applied to practice through, say, training, institutional incentives, organizational structures or public policies, the difficulty of improving outcomes is greatly amplified. Now the task involves moulding, coordinating and governing the activities of practitioners, who themselves must acquire judgement and skill that may not be easily translatable from one context to another. Interpreting the results of management or policy innovations is difficult because of the many variables involved, few of which are directly related to the actual technology deployment. When the results of applying knowledge to practice are uncertain, the value of the new knowledge itself becomes subject to controversy.

The limits of technology

In the absence of an existing standardized core, therefore, R&D programmes aimed at solving particular social problems should neither be expected to succeed, nor be advertised as having much promise of succeeding, at least in the short and medium term. They should be understood and described as aiming at the creation of fundamental knowledge and the exploration of new approaches, with success possible only over the long term, and with a significant chance of failure.

We are not, of course, arguing against working hard to address social problems that are not amenable to technological fixes, but we are saying that R&D is unlikely to be the main source of short- or medium-term progress. So long as the teaching of reading, for example, cannot be condensed into an easily deployed technology or tightly specified routine, improvement will mostly come through context-dependent trial and error at the level of public policy and organizational management. This is more a process of effective politics than effective innovation, and typically progress will be slow, hard-fought and uneven. On the other hand, when a standardized technological core relevant to a particular problem is available, appropriate R&D investments have the potential to stimulate rapid progress.

How might these insights help guide innovation policies today? To illustrate the implications of our rules for decision-makers, we turn briefly to climate change.

The global energy system that lies at the heart of the climate-change problem is probably more complex and resistant to successful reform than the health-care and education systems discussed above. Despite enormous scientific, political and diplomatic efforts over the past two decades, no progress on reducing global greenhouse-gas emissions has been made[5,6]. In the absence of technological fixes, progress towards significant reductions of greenhouse-gas emissions will remain frustratingly slow, uneven and inconclusive.

What are the prospects for a technological fix? In principle, stabilizing atmospheric carbon dioxide concentrations at levels deemed acceptable by climate experts can be achieved through radically reduced emissions or through direct removal of CO_2 from the atmosphere. Most discussion and effort focuses on the former. The suite of promising possibilities for reducing emissions — from nuclear fission, to photovoltaics, to on-site carbon capture and storage — offers attractive targets for R&D investments[7] consistent with Rule III: existing technological capacities can leverage continued improvement. Nevertheless, successful transition to a low-emissions energy system requires effective management across all sectors of society and all uses of CO_2-emitting technologies. Within this system, no particular technology fully encompasses the go of the process — eliminating CO_2 in the atmosphere — just as no particular teaching technology encompasses the go of teaching children to read. Rule I is violated.

> "Technological fixes do not offer a path to moral absolution, but to technical resolution."

Moreover, because emissions-reducing technologies will compete with existing energy technologies supported by entrenched interests, and because there will be competition between the emerging technologies, we can expect ongoing technical and political debates about efficacy of specific technologies, as seen for biofuels today — a violation of Rule II. System-wide progress is therefore likely to be buffered by political processes similar to the ones that frustrate progress now.

In contrast, direct removal of CO_2 from the atmosphere — air capture[8] — satisfies the rules for technological fixes. Most importantly, air capture embodies the essential cause–effect relations — the basic go — of the climate-change problem, by acting directly to reduce CO_2 concentrations, independent of the complexities of the global energy system (Rule I). There is a criterion of effectiveness that can be directly and unambiguously assessed: the amount of CO_2 removed (Rule II). And although air-capture technologies have been remarkably neglected in both R&D and policy discussions, they nevertheless seem technically feasible (Rule III)[9–11].

Our rules do not allow us to predict if air-capture technologies will in fact help stabilize greenhouse-gas concentrations. Certainly these technologies face technical, political and economic obstacles. Our rules do, however, allow us to strongly predict that stabilization is unlikely to be achieved, except in the very long term, without something like air capture. Such technologies should therefore receive much greater attention in energy innovation portfolios.

The climate-change example illustrates an important final point: technological fixes do not offer a path to moral absolution, but to technical resolution. Indeed, one of the key elements of a successful technological fix is that it helps to solve the problem while allowing people to maintain the diversity of values and interests that impede other paths to effective action. Recognizing when such opportunities for rapid progress are available should be a central part of innovation policy, and should guide investment choices. ■

Daniel Sarewitz is co-director of the Consortium for Science, Policy & Outcomes, and Professor of Science and Society, at Arizona State University, Tempe, Arizona 85287, USA.
e-mail: daniel.sarewitz@asu.edu
Richard Nelson is George Blumenthal Professor of International and Public Affairs Emeritus at Columbia University, New York 10027, USA, and visiting professor at the University of Manchester Business School.
e-mail: rrn2@columbia.edu

1. http://nces.ed.gov/nationsreportcard/ltt/results2004
2. Donovan, M. S., Wigdor, A. K. & Snow, C. E. (eds) *Strategic Education Research Partnership* (National Academies Press, 2003).
3. Foray, D. & Hargreaves, D. *Lond. Rev. Edu.* **1,** 7–19 (2003).
4. Murnane, R. & Nelson, R. *Economics of Innovation and New Technology* **16,** 307–327 (2007).
5. Prins, G. & Rayner, S. *Nature* **449,** 973–975 (2007).
6. Sarewitz, D. & Pielke Jr, R. in *Controversies in Science and Technology, vol. 2: From Chromosomes to the Cosmos* (eds Kleinman, D. L. & Handelsman, J.) 329–351 (Mary Ann Liebert, 2008).
7. Hoffert, M. I. *et al. Science* **298,** 981–986 (2002).
8. Keith, D. W., Ha-Duong, M. & Stolaroff, J. K. *Climatic Change* **74,** 17–45 (2006).
9. Baciocchi, R., Storti, G. & Mazzotti, M. *Chem. Eng. Process.* **45,** 1047–1058 (2006).
10. Lackner, K. S. *Science* **300,** 1677–1678 (2003).
11. Stolaroff, J. *Capturing CO_2 from Ambient Air: A Feasibility Assessment* (PhD Dissertation, Carnegie Mellon University, 2006).

To comment on this article and others in our innovation series, visit http://tinyurl.com/5uolx2.

Correction
The Commentary 'Towards responsible use of cognitive-enhancing drugs by the healthy' (*Nature* **456,** 702–705; 2008) described views derived from the report *Beyond Therapy* as solely those of Leon Kass. In fact, the work in question was by the President's Council on Bioethics, which at that time Kass chaired.

BOOKS & ARTS

Critical of quantum theory, Einstein (right) forced Bohr (left) and others to refine their thinking.

SCIENCE MUSEUM/SCIENCE & SOCIETY PICTURE LIBRARY

Entangled quantum histories

Two chronicles of quantum mechanics tell a good tale but don't reflect the conflicts between the physicists who struggled to reconcile theory and fact, explains Don Howard.

The nearly simultaneous publication of two popular histories of quantum mechanics, each highly readable and basically reliable, speaks well about the growing public interest in modern physics and about the commitment of major publishers to respond.

Louisa Gilder and Manjit Kumar take different approaches and emphasize different stages in the fascinating history of quantum physics. Kumar writes a conventional narrative history, focusing on the long-running debate between Niels Bohr and Albert Einstein, which took place from the mid-1920s through to the mid-1950s, over the adequacy of the quantum theory as a framework for fundamental physics. Gilder writes a delightfully unconventional history in the form of conversations — real or reconstructed — among the physicists themselves. She emphasizes the recent history of Bell's theorem, which concerns correlations between the quantum properties of separated elementary particles, its experimental tests and the subsequent exploitation of quantum entanglement in quantum computing, quantum information theory and quantum teleportation.

Gilder's is, on balance, the better book, partly because of the conversational format, which brings the scientist actors to life as complex personalities with interesting lives. Especially enjoyable are the portraits of the less famous physicists who, starting in the 1960s, put entanglement to the test and taught us how to engineer with it, starting with John Bell and including Abner Shimony, John Clauser, Alain Aspect and Anton Zeilinger.

The Age of Entanglement: When Quantum Physics Was Reborn
by Louisa Gilder
Knopf: 2008. 464 pp. $27.50

Quantum: Einstein, Bohr, and the Great Debate about the Nature of Reality
by Manjit Kumar
Icon Books: 2008. 480 pp. £20

Gilder has done her homework. Kumar relies uncritically on an old generation of specialist histories — such as Max Jammer's classic *The Conceptual Development of Quantum Mechanics* (McGraw-Hill; 1966), or the multivolume history of quantum physics (Springer; 1982–1987) by Jagdish Mehra and Helmut Rechenberg — but he does not recognize where those histories have been questioned or eclipsed by more recent work. Gilder, by contrast, has read the newer specialist literature. She has also done her own research, such as tracking down an old physics colleague of Boris Podolsky, and eliciting from him the striking memory that Podolsky said he and Nathan Rosen never asked Einstein for permission before putting his name on the classic 1935 Einstein, Podolsky and Rosen paper, 'Can Quantum Mechanical Description of Physical Reality Be Considered Complete?'.

One must, however, be cautious about reconstructing conversations. Even the participants' own memories are not always to be trusted, the conversations recalled in Werner Heisenberg's *Physics and Beyond* (Harper and Row; 1971) being a notorious case in point. And when an author gathers quotations from letters, memoirs or published papers covering a modest span of years and reassembles them as one exchange, there is as much fiction as fact in the product. Gilder's imagined contents of the lively chat between Bohr and Einstein that, time and again, kept them missing their tram stop during Einstein's Copenhagen visit in 1923 is wonderful as stagecraft. Yet I ask myself how soon I'll see a student presenting it in a term paper as if it were a verbatim transcript.

Two larger reservations must be recorded about both books. The first concerns their treatment of Bohr and what is commonly termed the Copenhagen interpretation of quantum mechanics. Both Kumar and Gilder take as fact much of the folklore and mythology surrounding the Copenhagen interpretation. Legend has it that, from 1927 onwards, the entire Copenhagen community spoke with one voice about questions of interpretation. In fact, from the beginning, Bohr and Heisenberg disagreed about many fundamental points. Bohr never endorsed such famous bits of alleged Copenhagen orthodoxy as wave-packet collapse, a privileged role for the subjective observer, or a distinction between quantum and classical levels of description based solely on a system's scale. Heisenberg, who invented the term Copenhagen interpretation much later in 1955, asserted all of these claims. Bohr did not. Differences between Heisenberg and Bohr go almost unnoticed by both Kumar and Gilder, with unfortunate consequences for such crucial issues as understanding what really was at stake in the Bohr–Einstein debate and the consequences for Bohr of experimental tests of Bell's theorem. These tests provide strong evidence for quantum entanglement — the entwined states of interacting quantum systems — that was the basis of Bohr's own interpretation of quantum mechanics.

The second reservation concerns the consequences of experimental tests of Bell's theorem for another major interpretive project — David Bohm's construction in 1952 of a 'non-local hidden variables theory'.

The theory posited unknown and perhaps inaccessible degrees of freedom of quantum systems, the values of which fix the behaviour of such systems. But quantum indeterminism is overcome at the price of allowing 'non-local' or 'faster-than-light' communication among those hidden variables, which might conflict with Einstein's theory of relativity. Gilder devotes long discussions to the career and ideas of Bohm. All the more puzzling, then, that when she tells us again and again about experiments confirming violations of the Bell inequality, she leaves the reader with the impression that this is a hands-down victory for quantum mechanics over hidden variable theories, when it is only local hidden variable theories that are put in jeopardy. One might reject Bohmian mechanics for other reasons, but these tests alone do not afford them. Kumar's hasty resumé of the same results leaves one still more confused. He suggests that Einstein's reservations about the completeness of quantum mechanics are somehow vindicated by experiments confirming the predictions of standard quantum mechanics.

Both authors would have been helped by a closer look at the well-known re-analysis of Bell's argument, introduced in 1984 by Jon Jarrett and refined by Shimony. Jarrett showed that the original Bell locality condition is a conjunction of two logically independent conditions that Shimony called outcome independence and parameter independence. The former is akin to a denial of quantum entanglement, the latter to relativistic locality constraints. Experimental violations of the Bell inequality can now be traced to violations of one or the other of these two conditions. That there are two independent routes to violations of the Bell inequality makes it clear how orthodox quantum mechanics, which presumes entanglement, and a Bohm-type hidden variables theory, which assumes relativistic non-locality at the microlevel, can both claim to have been vindicated by the Bell experiments. But these gentle criticisms should not deter the interested reader from enjoying two welcome additions to the popular history of twentieth-century physics. ■

Don Howard is professor of philosophy at the University of Notre Dame, Notre Dame, Indiana 46556, USA. His book on Einstein for Blackwell's *Great Minds* series will be published in 2009.
e-mail: dhoward1@nd.edu

Back to the roots of crop farming

Where Our Food Comes From: Retracing Nikolay Vavilov's Quest to End Famine
by Gary Paul Nabhan
Island Press: 2008. 266 pp. $24.95, £21.50

In 1941, when German and Finnish troops threatened to besiege the Russian city of Leningrad (now St Petersburg), Soviet leaders hurried to authorize the evacuation of the art collection from the city's Hermitage museum. Another extraordinary treasure, then the world's largest collection of more than 380,000 food crop samples housed at Leningrad's All-Union Institute of Agricultural Sciences, did not receive such privileged treatment; it survived the 1941–44 Leningrad blockade only through the virtue of committed individuals. Clearly, society places different values on the heritage content of the art museum and the seed bank.

In *Where Our Food Comes From*, Gary Paul Nabhan, a conservationist and research social scientist at the University of Arizona, spotlights crop diversity as a neglected but vital cultural resource. He does so by chronicling the journeys of Nikolay Vavilov, Russia's famous geneticist and botanist, and creator of the Leningrad seed bank. Vavilov was the first to identify the world's centres of crop diversity — a concept that, decades later, was developed by conservation biologists into what we now call biodiversity hotspots or conservation targets. Based on narratives from Vavilov's expeditions in the 1920s and 1930s and the author's own travels to retrace Vavilov's steps, Nabhan looks at changes in agricultural biodiversity.

The journeys cover nine regions of the globe. Each chapter highlights a specific land-use system, such as the wild apple forests of Kazakhstan; the diverse date cultivation in the oases of the Maghreb in north Africa; the traditional *milpa* cropping system in Mexico, based on maize (corn) and beans; or the forest gardens of the Amazon basin in South America. Nabhan observes that, almost everywhere, crop diversity has strongly and rapidly declined during the time between Vavilov's and his own visits. Both global- and regional-scale processes are held responsible for the homogenization of crops. For example, melting glaciers account for a depletion of hydrological sources in Tajikistan's Pamir mountains, which damages traditional cropping systems. In northern Italy's Po valley, wasteful irrigation practices and mechanized high-performance agriculture have displaced traditional land uses. In the Middle East, political unrest and the legacy of colonialism have disrupted ancient farming systems. In Kazakhstan, economic growth has driven urban sprawl, threatening agricultural land close to cities.

V. ENGLEBERT/TIME LIFE PICTURES/GETTY IMAGES

Amazon agriculture benefits from crop diversity.

Yet Nabhan's perspective is positive. He stresses the value of diverse crop varieties and the cultural traditions of land use and food preparation — for pragmatic reasons and as a cultural heritage that provides humankind with pleasure. He presents promising initiatives to recover crop diversity, from the rehabilitation of farmers' markets in Lebanon to efforts to renew local Hopi food systems in Arizona. And he shows how rich crop diversity and traditional seed selection and distribution support adaptation to changing environmental or societal conditions. I was fascinated by the story of how Colombian peasants have selected varieties of coca plants — the leaves of which contain cocaine alkaloids — that are resistant to the aerially sprayed herbicide used in the war on drugs. These resistant varieties have arisen through farmer-based selection, adaptation and informal dissemination of coca cuttings, all without any genetic engineering.

Where Our Food Comes From is a marked critique of the worldwide simplification of agricultural systems. It pins its hopes on local, traditional agriculture and is sceptical of top-down approaches to increasing food production, such as calls for another 'green revolution'. Many of its conclusions coincide with those reached by the International Assessment of Agricultural Knowledge, Science and Technology for Development, which concluded that the focus on maximum agricultural commodity production was responsible for the depletion of natural capital.

There are two opposing sides in the debate on how to increase food security. One side, which includes Nabhan, endorses the support of small-scale, low-output agriculture that is highly diverse. Among other examples, Nabhan uses the case of Ethiopian wheat varieties, which proved invaluable in the fight against

The theory posited unknown and perhaps inaccessible degrees of freedom of quantum systems, the values of which fix the behaviour of such systems. But quantum indeterminism is overcome at the price of allowing 'non-local' or 'faster-than-light' communication among those hidden variables, which might conflict with Einstein's theory of relativity. Gilder devotes long discussions to the career and ideas of Bohm. All the more puzzling, then, that when she tells us again and again about experiments confirming violations of the Bell inequality, she leaves the reader with the impression that this is a hands-down victory for quantum mechanics over hidden variable theories, when it is only local hidden variable theories that are put in jeopardy. One might reject Bohmian mechanics for other reasons, but these tests alone do not afford them. Kumar's hasty resumé of the same results leaves one still more confused. He suggests that Einstein's reservations about the completeness of quantum mechanics are somehow vindicated by experiments confirming the predictions of standard quantum mechanics.

Both authors would have been helped by a closer look at the well-known re-analysis of Bell's argument, introduced in 1984 by Jon Jarrett and refined by Shimony. Jarrett showed that the original Bell locality condition is a conjunction of two logically independent conditions that Shimony called outcome independence and parameter independence. The former is akin to a denial of quantum entanglement, the latter to relativistic locality constraints. Experimental violations of the Bell inequality can now be traced to violations of one or the other of these two conditions. That there are two independent routes to violations of the Bell inequality makes it clear how orthodox quantum mechanics, which presumes entanglement, and a Bohm-type hidden variables theory, which assumes relativistic non-locality at the microlevel, can both claim to have been vindicated by the Bell experiments. But these gentle criticisms should not deter the interested reader from enjoying two welcome additions to the popular history of twentieth-century physics. ■

Don Howard is professor of philosophy at the University of Notre Dame, Notre Dame, Indiana 46556, USA. His book on Einstein for Blackwell's *Great Minds* series will be published in 2009.
e-mail: dhoward1@nd.edu

Back to the roots of crop farming

Where Our Food Comes From: Retracing Nikolay Vavilov's Quest to End Famine
by Gary Paul Nabhan
Island Press: 2008. 266 pp. $24.95, £21.50

In 1941, when German and Finnish troops threatened to besiege the Russian city of Leningrad (now St Petersburg), Soviet leaders hurried to authorize the evacuation of the art collection from the city's Hermitage museum. Another extraordinary treasure, then the world's largest collection of more than 380,000 food crop samples housed at Leningrad's All-Union Institute of Agricultural Sciences, did not receive such privileged treatment; it survived the 1941–44 Leningrad blockade only through the virtue of committed individuals. Clearly, society places different values on the heritage content of the art museum and the seed bank.

In *Where Our Food Comes From*, Gary Paul Nabhan, a conservationist and research social scientist at the University of Arizona, spotlights crop diversity as a neglected but vital cultural resource. He does so by chronicling the journeys of Nikolay Vavilov, Russia's famous geneticist and botanist, and creator of the Leningrad seed bank. Vavilov was the first to identify the world's centres of crop diversity — a concept that, decades later, was developed by conservation biologists into what we now call biodiversity hotspots or conservation targets. Based on narratives from Vavilov's expeditions in the 1920s and 1930s and the author's own travels to retrace Vavilov's steps, Nabhan looks at changes in agricultural biodiversity.

The journeys cover nine regions of the globe. Each chapter highlights a specific land-use system, such as the wild apple forests of Kazakhstan; the diverse date cultivation in the oases of the Maghreb in north Africa; the traditional *milpa* cropping system in Mexico, based on maize (corn) and beans; or the forest gardens of the Amazon basin in South America. Nabhan observes that, almost everywhere, crop diversity has strongly and rapidly declined during the time between Vavilov's and his own visits. Both global- and regional-scale processes are held responsible for the homogenization of crops. For example, melting glaciers account for a depletion of hydrological sources in Tajikistan's Pamir mountains, which damages traditional cropping systems. In northern Italy's Po valley, wasteful irrigation practices and mechanized high-performance agriculture have displaced traditional land uses. In the Middle East, political unrest and the legacy of colonialism have disrupted ancient farming systems. In Kazakhstan, economic growth has driven urban sprawl, threatening agricultural land close to cities.

Amazon agriculture benefits from crop diversity.

V. ENGLEBERT/TIME LIFE PICTURES/GETTY IMAGES

Yet Nabhan's perspective is positive. He stresses the value of diverse crop varieties and the cultural traditions of land use and food preparation — for pragmatic reasons and as a cultural heritage that provides humankind with pleasure. He presents promising initiatives to recover crop diversity, from the rehabilitation of farmers' markets in Lebanon to efforts to renew local Hopi food systems in Arizona. And he shows how rich crop diversity and traditional seed selection and distribution support adaptation to changing environmental or societal conditions. I was fascinated by the story of how Colombian peasants have selected varieties of coca plants — the leaves of which contain cocaine alkaloids — that are resistant to the aerially sprayed herbicide used in the war on drugs. These resistant varieties have arisen through farmer-based selection, adaptation and informal dissemination of coca cuttings, all without any genetic engineering.

Where Our Food Comes From is a marked critique of the worldwide simplification of agricultural systems. It pins its hopes on local, traditional agriculture and is sceptical of top-down approaches to increasing food production, such as calls for another 'green revolution'. Many of its conclusions coincide with those reached by the International Assessment of Agricultural Knowledge, Science and Technology for Development, which concluded that the focus on maximum agricultural commodity production was responsible for the depletion of natural capital.

There are two opposing sides in the debate on how to increase food security. One side, which includes Nabhan, endorses the support of small-scale, low-output agriculture that is highly diverse. Among other examples, Nabhan uses the case of Ethiopian wheat varieties, which proved invaluable in the fight against

the virulent leaf rust fungus, to show that traditional, allegedly outmoded forms of land use may help to solve modern environmental problems. The opposing side advocates modern, technology-intensive, high-performance forms of agriculture that safeguard large parts of the world's food supply today, but are not resilient and depend on significant inputs of fossil fuels that may become expensive in the future.

Nabhan notes the often overlooked contributions of traditional ecological knowledge to sustainable food production. After reading the book, major questions arise. Do locally domesticated plants have an archival function that provides genetic varieties for modern plant breeding, and should they thus be conserved in their remote centres of origin? Or can they be used on a larger scale in intensive agricultural landscapes? The integration of traditional and modern practices into agricultural systems that are productive yet sustainable, consideration of the needs of small-scale farmers and maintenance of diverse ecosystems will all remain conflict-laden — but they represent crucial challenges in the quest to feed the world. ■

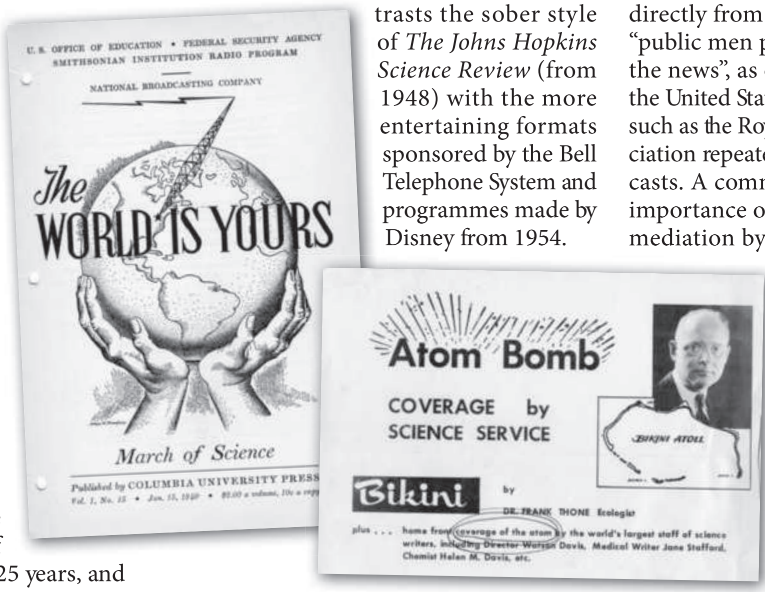
Tobias Plieninger heads a group on ecosystem services at the Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstrasse 22/23, 10117 Berlin, Germany.
e-mail: plieninger@bbaw.de

Early days of science broadcasting

Science on the Air: Popularizers and Personalities on Radio and Early Television
by Marcel Chotkowski LaFollette
Chicago Univ. Press: 2008. 324 pp. $27.50

Writing popular science and writing about popular science are very different endeavours. The former is at least as old as John Newbery's *Newtonian System of Philosophy* (1761), with its star Tom Telescope, and has been practised in every medium in turn: lectures, books, newspapers, magazines, cinema, radio, television and the Internet. But analysing it is a recent pursuit, tied to the growth of university science-studies departments concerned with science's social dimensions. Although now established, the field is not so overpopulated that there is agreement about how to do it, or for whom.

Marcel Chotkowski LaFollette has written on the history of popular science for more than 25 years, and is best known for *Making Science our Own: Public Images of Science, 1910–1955* (Univ. Chicago Press, 1990). In her new book, *Science on the Air*, she reveals from primary sources the story of the first 20 years of science broadcasting on US radio. We meet the Smithsonian Institution's Austin Hobart Clark, who initiated a series of scientists' radio lectures in 1923, and Thornton Burgess, who transferred from writing children's nature books to radio programmes. LaFollette discusses how the scientist William Ritter and the newspaper publisher Edward Scripps established the Science Service news bureau in 1920, which was highly influential in both print and radio for four decades. The one post-war chapter, which benefits from an established secondary literature, discusses science on US television in the 1940s and 1950s. Here LaFollette contrasts the sober style of *The Johns Hopkins Science Review* (from 1948) with the more entertaining formats sponsored by the Bell Telephone System and programmes made by Disney from 1954.



Scientists conveyed information directly to the public in early radio broadcasts.

SMITHSONIAN INSTITUTION ARCHIVES

Science on the Air has an academic level of detail. But LaFollette writes in a narrative style, with occasional touches of whimsy — "microphones trembled during atomic bomb tests and rattled with impassioned debate", for example. This choice has consequences; after reading it, you will know what happened to US science radio and early television, but you may wonder why you need to know it. A more argumentative or analytical approach would have addressed why popular science matters.

The relationship of science to people and to the broader culture was, in the period she describes, gaining importance as the sciences began to have a significant impact on ordinary lives. The possibility that scientists might have a negative impact, through their perceived responsibility for new weapons and mass-production machinery, was a live issue. More historical studies of the vehicles of science communication are needed so that we can understand the interplay between science and society; but this requires more than narrative.

So little is published in this area that it would be unfair to ask LaFollette to draw international comparisons. With a myriad of local commercial stations to follow and many potential stories to tell, she has a more difficult job than scholars of British science radio, for example, where for the same period there was one monopoly broadcaster, the BBC. LaFollette describes how in the United States scientists preferred a lecture format, but were put under pressure to make their programmes more entertaining or risk marginalization. In Britain, producers remained deferential and accepted that information should be conveyed directly from the authority to the viewer — "public men personally discussing events in the news", as one producer put it. Unlike in the United States, elite scientific organizations such as the Royal Society and the British Association repeatedly sought control over broadcasts. A common element was the growing importance of the skilled intermediary, and mediation by a presenter was the norm on television by the late 1950s.

Today the media market is global. Producers work in a multichannel world, where no country's output can afford to limit itself to national concerns. Single broadcasters can rarely afford the cost of ambitious science shows and so seek to collaborate. Co-production has led to a homogenization of style, moving closer to the US entertainment model and away from the earnest, deferential treatment LaFollette describes in 1920s radio. Some commentators fear that the science in science programming is reaching homeopathic concentration. The continuing presence of science in the schedules, evidently vulnerable to broadcasting fashions, demands investigation. This book shows it can be done. ■

Tim Boon is chief curator at the Science Museum, London, and author of *Films of Fact: A History of Science in Documentary Films and Television*.

Trading Bronze Age technology

Beyond Babylon: Art, Trade, and Diplomacy in the Second Millennium B.C.
Metropolitan Museum of Art, New York
Until 15 March 2009

Before the BlackBerry, before the iPhone, before e-mailing, texting and instant messaging, was an ivory-hinged boxwood writing board. Originally inlaid with wax, it was probably engraved with a ship's cargo inventory. The hand-sized board, now shrivelled and cracked, might also have carried messages between ancient Near Eastern courts, and sank to the bottom of the Mediterranean Sea at the end of the fourteenth century bc with its ship: a 15-metre-long vessel wrecked off the coast of southern Turkey.

The ship's contents, currently on display in New York in the Metropolitan Museum of Art's dazzling exhibition, *Beyond Babylon*, are a testament to the globalization of trade and mass communication that existed in the Bronze Age. The earliest seagoing vessel ever excavated, the ship carried 15,000 items from 12 different cultures, including Old Babylonian, Egyptian, Mycenaean, Assyrian, Canaanite, Nubian, Balkan and Baltic. The 17 tonnes of cargo included 10 tonnes of copper from Cyprus; 1 tonne of tin, probably mined in central Asia; 350 kilograms of turquoise and purple glass ingots from Egypt; ebony from Nubia (now northern Sudan); and 150 Canaanite jars, many filled with resin from the terebinth tree, used in scented oils and as an antiseptic. Precious gifts included ostrich eggs, gold pendants, glass beads, hippopotamus teeth, spices such as coriander and cumin, and two delicately carved, duck-shaped ivory pots.

T. FARRAJ

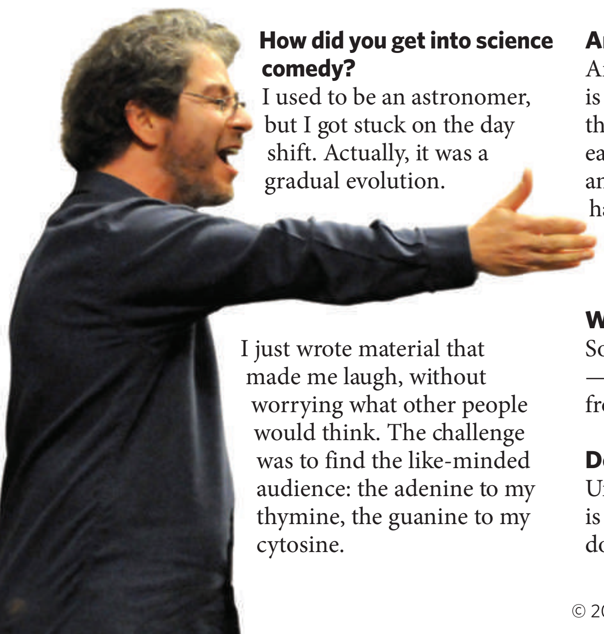
The quest for commodities such as copper and tin, the basic ingredients of bronze — and for exotic materials such as ivory and the semiprecious stone lapis lazuli — drove commerce between these kingdoms of the second millennium bc, and inspired a cross-border blossoming of art and technology. The influence of one culture on another is evident in the mélange of styles incorporated in the artefacts on show, which encompass a sumptuous array of gold diadems and scarabs, jewel-inlaid daggers, bronze deities (detail, pictured), carved seals and ceramic jars from sites stretching from the Mediterranean to the Indus River. Another duck-shaped container, from the reign of Egypt's Amenhotep III (around 1390 to 1352 bc), has carved ivory wings of Canaanite design, but the duck is held in the outstretched arms of a lissom female swimmer with Egyptian features. A silver stag-shaped cup, from the Hittite empire of Anatolia in the fourteenth century bc, incorporates hunt scenes and is inlaid with a copper alloy used earlier in Middle Kingdom Egypt. The transfer of technology is also apparent in the blue glass ingots excavated from the shipwreck: coloured with cobalt, they are chemically similar to vessels made of blue glass from Amarna, Egypt, and to blue beads from Mycenaean Greece.

Artisans and scholars also moved across cultures. A fourteenth-century bc fragment of a cuneiform clay tablet from the ancient Hittite capital Boğazkale, now in modern Turkey, is carved with the Epic of Gilgamesh, the Mesopotamian saga that visiting professors circulated far and wide. Equally fascinating is 'The Game of 58 Holes', a board game similar to snakes and ladders that was found across Mesopotamia, Iran and Anatolia. The exhibition includes an ebony-and-ivory version etched with a palm tree and studded with ten dog-headed pegs; it was found in Egyptian Thebes and dates to 1981 to 1802 bc. It may not be an iPod, but even then, perhaps, tired traders on long trips found entertaining ways to break the monotony. ■

Josie Glausiusz is a journalist based in New York.
e-mail: josiegz@earthlink.net

Q&A: Helium walks into a bar…

Science comedian **Brian Malow** is a regular performer on the museum and conference circuit in the United States. He explains why he finds science funny, and how he uses comedy to gain the public's interest.



J. GILBEY

How did you get into science comedy?
I used to be an astronomer, but I got stuck on the day shift. Actually, it was a gradual evolution. I just wrote material that made me laugh, without worrying what other people would think. The challenge was to find the like-minded audience: the adenine to my thymine, the guanine to my cytosine.

Are some topics funnier than others?
Anything that you can anthropomorphize is easy — such as insects and animals. And the more familiar I am with a subject, the easier it is. That's why I have a lot of physics and astronomy material. The other day, I had a frightening experience at a café when I noticed someone had put pasta and antipasto right next to each other.

What do you do when a joke fails?
Sometimes I say, "That joke was endothermic — it required the addition of a little energy from you to make it work."

Do science jokes go out of fashion?
Unlike politics or pop culture, basic science is evergreen. Gravity and laws of motion don't go out of style. Then there's always new science to cover, such as the Large Hadron Collider. It has already proven science can get massive media coverage if it's rumoured it might destroy the planet.

Is scientific jargon a source of humour?
Absolutely. I did some shows for the American Chemical Society and prepared by studying a glossary of chemical terms. There's a lot of pun potential in chemistry. They wondered if I could be off-the-cuff and I said, "I'm so spontaneous I have a negative ΔG."

And that helium?
Right. So the bartender says, "We don't serve noble gases here." The helium doesn't react. ■

Interview by **Nick Thomas**, associate professor of chemistry at Auburn University at Montgomery, Alabama 36117, USA.
e-mail: nthomas@aum.edu

Brian Malow: The Final Frontier?
Next show at the National Institute of Standards and Technology, Maryland
19 December 2008
www.sciencecomedian.com

ORGANIC CHEMISTRY

How to beat an alcohol problem

Karl B. Hansen

Some alcohols can't be made in a way that controls the crucial arrangement of chemical groups in the molecule. A reaction that uses common laboratory reagents offers a practical route to these compounds.

Since the discovery that many organic molecules have mirror-image forms[1], chemists have sought to control which of these 'enantiomers' is synthesized in reactions. Several strategies have been developed, such as the use of catalysts[2] or enzymes[3] to direct reactions towards making a particular chiral form. Alternatively, desired enantiomers can be separated from mixtures using chromatography[4] or other 'resolution' techniques[5]. Certain classes of compounds can now be easily prepared as pure enantiomers, but others — in particular those with many chemical groups attached to a single carbon atom — remain a challenge.

Reporting on page 778 of this issue, Aggarwal and colleagues[6] describe a striking reaction in which an alkyl (hydrocarbon) group is attached to a chiral carbon atom so that the enantiomeric purity of the reactant is retained in the product. Because a compound's enantiomers often have different biological activities, this could be particularly useful for drug discovery and development.

So what is the molecular basis of chirality in molecules? When four different chemical groups (or atoms) are attached to a carbon atom, two geometrically distinct arrangements of the groups around the carbon are possible (Fig. 1a). The two resulting molecules are mirror images, and can't be superimposed on each other. These molecules behave identically in chemical reactions unless they interact with other chiral molecules — in the same way that a left hand is, to all intents and purposes, the same as a right hand until you try to insert it into a right-handed glove.

Because almost all of the components in cells — proteins, nucleic acids and so on — are chiral, two enantiomers of the same molecule will almost always interact with those components in different ways. An everyday example of this is a molecule called carvone[7]: one enantiomer smells of caraway, whereas the other smells of spearmint, because of the way each isomer interacts with olfactory receptors.

In the case of chiral medicines, the consequences of such isomerism can be grave. One enantiomer of a drug might bind to its intended target protein much more strongly than the other. Even worse, the 'wrong' form

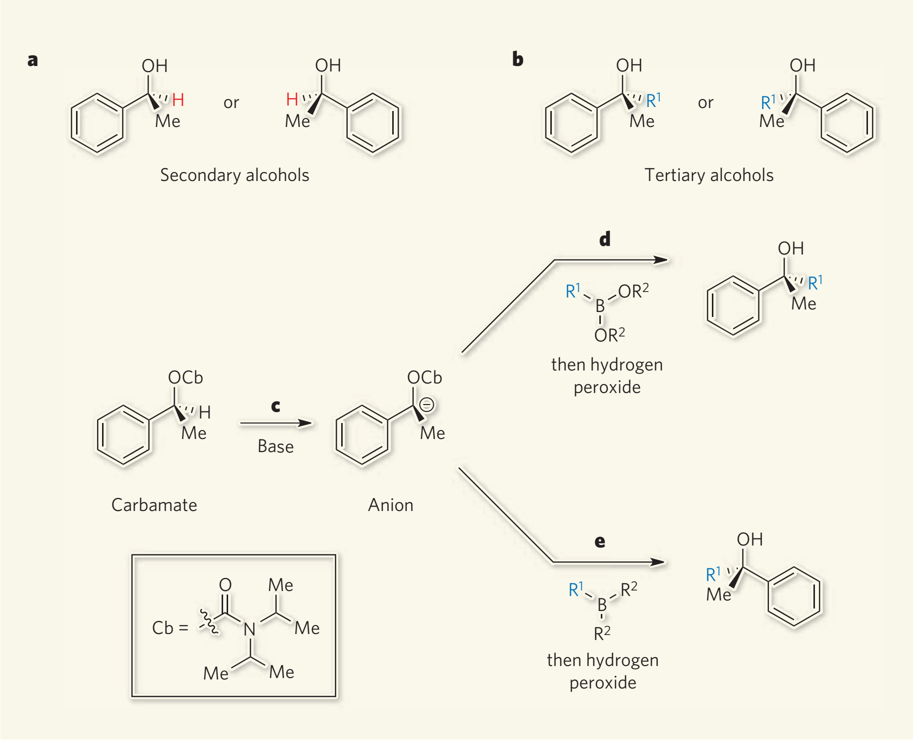


Figure 1 | Preparing single enantiomers of tertiary alcohols. **a**, Organic compounds that contain carbon atoms to which four different groups are attached can form as two possible enantiomers — mirror images of each other, as shown for this secondary alcohol (Me is a methyl group). Secondary alcohols can be easily synthesized as single enantiomers. **b**, Tertiary alcohols are structurally similar to secondary alcohols, but contain an extra hydrocarbon group (in this case, hydrocarbon R^1 replaces the hydrogen atoms shown in red in **a**). Single enantiomers of tertiary alcohols cannot be made using methods analogous to those used for secondary alcohols. Aggarwal and colleagues[6] report a method that solves this problem (**c**–**e**). **c**, Carbamates (readily made from secondary alcohols) can be treated with a base to generate an anion. **d**, If the anion is reacted with a boronate ester ($R^1B(OR^2)_2$, where R^2 is any hydrocarbon group) and then with hydrogen peroxide, a single enantiomer of a tertiary alcohol is produced, which retains the arrangement of groups in the carbamate. **e**, If a borane ($R^1BR^2_2$) is used instead of a boronate ester, the opposite enantiomer is made.

might interact with a completely different receptor. The most drastic and oft-cited case of this is thalidomide[8]: one enantiomer is a sedative, the other causes severe birth defects. Drugs for human use were often sold as one-to-one mixtures of enantiomers, because such mixtures are easier to prepare than the single isomers. But since the thalidomide tragedy, drug-regulatory agencies strongly prefer chiral molecules to be prepared and marketed as pure enantiomers[9].

Chemists have risen to this synthetic challenge magnificently. It is now possible to prepare many kinds of compound routinely in enantiomerically pure form, including a structural class known as secondary alcohols. Secondary alcohols are useful as chiral building blocks for more complex, biologically active molecules, and often have interesting biological activities in their own right. They consist of a hydrogen atom, a hydroxyl (OH) group and two alkyl groups attached to a central

carbon atom (Fig. 1a). One of the most successful ways to prepare single enantiomers of secondary alcohols is to add hydrogen to carbonyl compounds (which contain carbon–oxygen double bonds) in the presence of a chiral catalyst.

But alcohols of another class are much more difficult to make as pure enantiomers. Tertiary alcohols contain carbon atoms to which a hydroxyl group and three alkyl groups are attached (Fig. 1b), and are just as useful as secondary alcohols. In principle, they could be prepared by adding an alkyl group to a carbonyl compound in the presence of a chiral catalyst. Unfortunately, although chemists have developed many powerful catalysts that allow pure enantiomers of secondary alcohols to be made by adding hydrogen to carbonyls, there are far fewer general methods that control the addition of alkyl groups to carbonyls to make single enantiomers of tertiary alcohols[10].

Aggarwal and colleagues[6] have therefore taken a different approach. They have devised a way to transform chiral molecules known as carbamates — which can be prepared easily from secondary alcohols — into tertiary alcohols (Fig. 1c–e). The method involves plucking a hydrogen atom from the carbamate using a base, thus forming an intermediate anion. The anion is then reacted with a boron-containing reagent, which acts as a source of alkyl groups. Finally, an oxidizing agent (hydrogen peroxide) is added. This acts as a source of hydroxyl groups and its addition results in the formation of the desired tertiary alcohol.

There are several interesting and useful aspects of this work. First, the chiral 'information' in the starting material is transferred efficiently to the product — that is, an enantiomerically pure reactant is converted into essentially a single enantiomer of product. Often, the anions of chiral compounds quickly convert back and forth between enantiomers, so that the chiral information from the reactant is lost and the product forms as a mixture of isomers. But Aggarwal and colleagues form their anion intermediates under conditions in which this interconversion is slow, thus preventing the formation of a mixture.

Another impressive feature of the method is that it allows many different tertiary alcohols to be prepared from a single carbamate. A wide array of boron reagents is commercially available (or can be easily prepared) that can act as sources of different alkyl groups. This means that a diverse set of products can be made conveniently from a common starting material.

The most remarkable aspect of this process, however, is that the exact enantiomer of the product prepared in the reaction is governed by the nature of the boron reagent selected. The authors observed that the use of a boronate ester (which contains two boron–oxygen bonds and one boron–carbon bond, Fig. 1d) results in the groups around the chiral carbon atoms retaining the same arrangement in the product as in the carbamate. But when they used a borane (which contains three boron–carbon bonds and no boron–oxygen bonds, Fig. 1e), the groups 'flip' into the mirror-image arrangement. Thus, either enantiomer of the product can be prepared from a single chiral precursor — a rare feature that is not only practically useful, but also interesting in terms of fundamental reactivity.

Aggarwal and colleagues' findings[6] are likely to inspire others to explore the chemistry of stable chiral anions in organic molecules, opening up new areas of research. Moreover, by replacing hydrogen peroxide in the final step of the procedure with other reagents, it might be possible to make other chiral molecules that have been difficult to prepare as single enantiomers. But what is particularly exciting is how rapidly this work can be put into practice by others in the field: many of the required reagents and substrates are common items in the stockrooms of organic chemistry labs around the world. ■

Karl B. Hansen is in the Department of Small Molecule Process and Product Development, Amgen, One Kendall Square, Building 1000, Cambridge, Massachusetts 02139, USA.
e-mail: karl@amgen.com

1. Eliel, E. L. & Wilen, S. H. *Stereochemistry of Organic Compounds* (Wiley, 1994).
2. Walsh, P. J. & Kozlowski, M. C. *Fundamentals of Asymmetric Catalysis* (Univ. Sci. Books, 2008).
3. *Biocatalysis* (*Nature* Insight) *Nature* **409,** 225–268 (2001).
4. Cox, G. B. (ed.) *Preparative Enantioselective Chromatography* (Blackwell, 2005).
5. Jacques, J., Collet, A. & Wilen, S. H. *Enantiomers, Racemates and Resolutions* (Wiley, 1981).
6. Stymiest, J. L., Bagutski, V., French, R. M. & Aggarwal, V. K. *Nature* **456,** 778–782 (2008).
7. Leitereg, T. J., Guadangi, D. G., Harris, J., Mon, T. R. & Teranishi, R. *J. Agric. Food Chem.* **19,** 785–787 (1971).
8. Rouhi, M. *Chem. Eng. News* **83** (25), 122–123 (2005).
9. www.fda.gov/cder/guidance/stereo.htm
10. Riant, O. & Hannedouche, J. *Org. Biomol. Chem.* **5,** 873–888 (2007).

MICROBIAL ECOLOGY

Metabolism of the deep

Christa Schleper

Certain microorganisms from the domain Archaea seem to be big players in the marine carbon and nitrogen cycles. A study linking their abundance in the deep sea to their likely metabolic profile refines this view.

The deep ocean covers two-thirds of our planet and teems with microbial life. Understanding the roles of deep-sea microbial communities is therefore essential for understanding global biogeochemical cycling, which in turn is pivotal for all other forms of life. Some of the most abundant creatures in the deep sea are the Crenarchaeota, microorganisms that belong to the domain Archaea. In recent years, it has been suggested[1–4] that members of one major group of Crenarchaeota have a central role in the ocean's nitrogen cycle, because they perform the first step in the process that converts ammonia to nitrate. Furthermore, they were thought to be autotrophs (organisms that fix inorganic carbon for growth, rather than acquiring carbon from organic sources), thus also implicating them in the carbon cycle.

But on page 788 of this issue[5], Agogué *et al.* propose that neither of these ideas is true for all Crenarchaeota. Their data suggest that some of these organisms — especially those residing in the deeper oceanic waters of the subtropical and tropical North Atlantic Ocean — do not carry genes for the main enzyme involved in ammonia oxidation, and that deep-sea Crenarchaeota depend on organic sources of carbon.

Agogué and colleagues' findings contribute to the current debate about the ecological role of Crenarchaeota in marine and terrestrial environments. Variants of genes that encode ammonia monooxygenase — the enzyme that oxidizes ammonia (Fig. 1) — were recently detected[1,2,4,6] in archaeal genomes. When coupled with the fact that there are hundreds (and sometimes thousands) more archaea than ammonia-oxidizing bacteria in many terrestrial and marine environments, archaea suddenly emerged as probable major contributors to the nitrogen cycle[2,3,5,7,8].

This finding was surprising because, for the past hundred years, ammonia oxidation had been considered to be performed exclusively by specialist bacteria. The isolation and laboratory cultivation of ammonia-oxidizing archaea confirmed the hypothesis that certain Crenarchaeota are capable of generating energy through ammonia oxidation in an oxygen-dependent process, and thereby fix carbon from inorganic sources, just like their bacterial counterparts[1,9,10]. But it remained to be shown whether the high abundance of the archaeal *amoA* gene (which encodes ammonia monooxygenase) in the environment reflects a high archaeal contribution to ammonia oxidation. Furthermore, there was the question of whether all Crenarchaeota in moderate environments contribute to ammonia oxidation.

Apparently not. Agogué *et al.*[5] do not find evidence for ammonia oxidizers in the deep sea of equatorial regions. They took water samples from different depths in the North Atlantic Ocean, and at latitudes ranging from polar

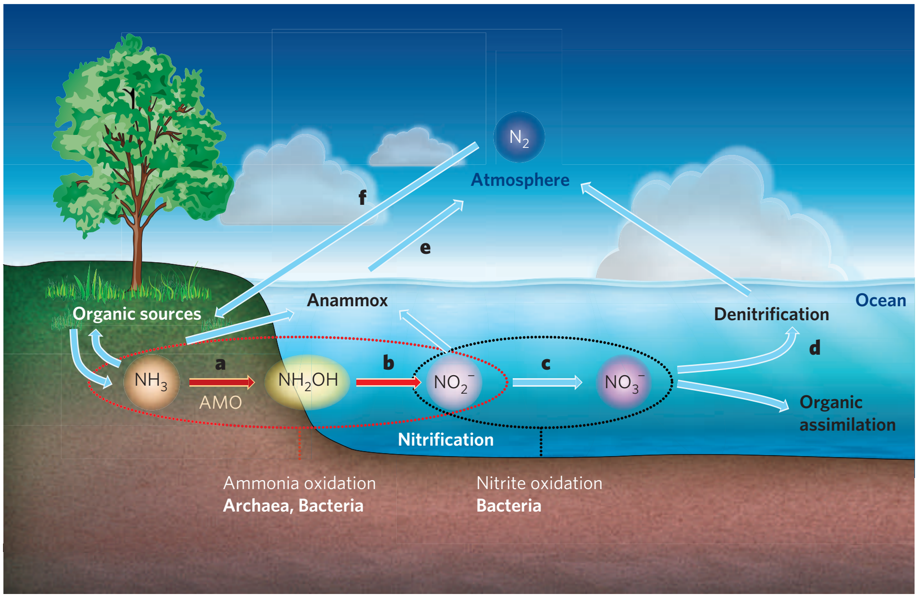


Figure 1 | Nitrification in the global nitrogen cycle. Ammonia oxidation is the first step of nitrification, the process in which ammonia (NH_3, of organic origin) is oxidized to nitrate ions (NO_3^-). Nitrification occurs on land and in the oceans and is essential for the global cycling of nitrogen. The process was long thought to be performed solely by bacteria, but certain archaea (Crenarchaeota) have recently been shown to be capable of ammonia oxidation, and to contain genes for ammonia monooxygenase (AMO, the key enzyme involved in ammonia oxidation). **a**, AMO converts ammonia into hydroxylamine (NH_2OH). **b**, Hydroxylamine is converted by the same microorganisms into nitrite ions (NO_2^-) — a protein that catalyses this process has not yet been found in archaea. **c**, Other specialized bacteria complete nitrification by converting nitrite ions to nitrate ions. **d**, The nitrate is then either assimilated into organic matter or denitrified by other microorganisms to produce nitrogen, which escapes into the atmosphere. **e**, 'Anammox' bacteria can also convert ammonia and nitrite ions into nitrogen. **f**, Nitrogen is fixed mostly by specialized bacteria, producing ammonia. This can be incorporated into organic matter, or oxidized as the cycle continues. Agogué *et al.*[5] show that ammonia-oxidizing archaea are abundant in the North Atlantic Ocean, but not in the deep sea of equatorial regions.

regions down to the Equator. They then quantified the ratio of the number of archaeal *amo*A genes in the samples to the total number of Crenarchaeota, the latter number being determined by measuring the abundance of an RNA gene that is ubiquitous in these organisms.

The results[5] indicate that the ratio decreases markedly from subpolar to equatorial regions, and also from subsurface waters to the deep ocean. Whereas the concentrations of *amo*A genes found at all depths in polar regions were high, they were 1,000 times lower than the concentration of total Crenarchaeota found in the bathypelagic waters (1,000–4,000 metres) of subtropical regions. The authors therefore conclude that most deep-sea Crenarchaeota in subtropical waters do not oxidize ammonia. Perhaps tellingly, the concentration of ammonia in subtropical deep waters is considerably lower than in the equivalent polar regions, where potential ammonia oxidizers are found in high numbers.

Agogué *et al.* also measured the ability of microorganisms taken from the locations described above to fix inorganic carbon. They found that the decreases in *amo*A abundance seen in their experiments[5] mirror the observed patterns of carbon fixation — as the ability of marine microorganisms to fix carbon declines, so does the number of *amo*A genes in the archaeal population. This in turn suggests that deep-sea Crenarchaeota are not autotrophs.

Although these findings[5] significantly refine our perception of crenarchaeotal metabolisms in the ocean, some caveats remain. The authors' gene-detection method assumes that archaeal *amo*A genes are similar to known *amo*A genes. But it is difficult to rule out the possibility that Agogué *et al.* simply missed specific variations of *amo*A genes that are typical of deep-water Crenarchaeota alone. Numerous cases exist in microbial ecology in which DNA-detection methods completely overlooked specific groups of genes, including *amo*A genes.

Similarly, Agogué and colleagues' data do not conclusively prove that deep-sea Crenarchaeota are not autotrophic. But their conclusions are supported by evidence from a series of previous studies[11–14] that reported the incorporation of organic carbon sources into Crenarchaeota, thus indicating that at least some Crenarchaeota are heterotrophic — they take up organic carbon for growth.

It was not clear from the earlier studies[11–14] whether heterotrophic growth is a fallback option for otherwise autotrophic, ammonia-oxidizing archaea, or whether distinct populations of Crenarchaeota exist that are strictly either heterotrophic or autotrophic. Agogué and colleagues' findings[5] suggest the existence of specific heterotrophic Crenarchaeota,

50 & 100 YEARS AGO

50 YEARS AGO

The first symposium on space medicine to be held in Great Britain was held at B.M.A. House, London, on October 16 and 17 ... The symposium was arranged to discuss the reactions of man to the conditions that will be encountered in space and on other planets, and of the equipment that will be required for survival in these hostile environments ... It is certain that man will travel into space within a few years. His first venture will probably be in orbit, close to the Earth, and lasting a few hours. His second venture will be to the Moon, and his third to one of the planets. Voyages beyond our own solar system are difficult to comprehend, but should not be dismissed too lightly. The reason for man travelling into space is a controversial subject. Some contend that he merely adds to the complexity of the space vehicle; others that he will earn his keep by taking decisions ... The fact remains, however, that he will go, and that he will be faced with many of the problems discussed at the symposium.

From *Nature* 13 December 1958.

100 YEARS AGO

From Prof. Haberlandt's laboratory at Graz there has been issued another paper on the perception of light by plants ... The author, Dr. K. Gaulhofer, has studied the epidermal cells of the leaves of certain plants that take up a fixed light position, and suggests that an explanation may be found in the presence of pits or clefts in the cell walls acting as light distributors ... The rays of light impinging on the edges of the pits are deflected, and consequently, underneath the pits, shadows are produced. Good instances of such pits occur in *Aporrhiza paniculata* and *Banisteria splendens*, while *Hyperbaena laurifolia* and *Abuta concolor* show well-marked clefts. The combination of pits and curved cell wall in *Cocculus laurifolius* will repay examination.

From *Nature* 10 December 1908.

devoid of the *amoA* gene, that live preferably in deep ocean waters (see Suppl. Fig. S4 of Agogué *et al.*).

Nevertheless, the success of archaea in many disparate environments might still turn out to be based on the organisms' ability to switch from an autotrophic to a heterotrophic lifestyle, depending on environmental conditions. Several lines of evidence suggest that all (or most) soil Crenarchaeota are potential ammonia oxidizers[7,8]. Yet it is difficult to imagine that the large populations of these archaea in terrestrial habitats are merely based on this form of metabolism, especially where ammonia and/or oxygen concentrations are low. So although molecular studies such as those of Agogué *et al.* are undoubtedly valuable in raising hypotheses, additional techniques — including methods for probing isotope uptake by single cells, techniques for cultivating microorganisms in the laboratory, and biochemical analyses — will be indispensable for shedding light on the physiology and ecological impact of the ubiquitous and enigmatic Crenarchaeota. ■

Christa Schleper is in the Department of Genetics in Ecology, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.
e-mail: christa.schleper@univie.ac.at

1. Könneke, M. *et al. Nature* **437,** 543–546 (2005).
2. Schleper, C., Jurgens, G. & Jonuscheit, M. *Nature Rev. Microbiol.* **3,** 479–488 (2005).
3. Wuchter, C. C. *et al. Proc. Natl Acad. Sci. USA* **103,** 12317–12322 (2006).
4. Hallam, S. J. *et al. PLoS Biol.* **4,** e95 (2006).
5. Agogué, H., Brink, M., Dinasquet, J. & Herndl, G. J. *Nature* **456,** 788–791 (2008).
6. Francis, C. A. *et al. Proc. Natl Acad. Sci. USA* **102,** 14683–14688 (2005).
7. Leininger, S. D. *et al. Nature* **442,** 806–809 (2006).
8. Prosser, J. I. & Nicol, G. W. *Environ. Microbiol.* **10,** 2931–2941 (2008).
9. de la Torre, J. R., Walker, C. B., Ingalls, A. E., Könneke, M. & Stahl, D. A. *Environ. Microbiol.* **10,** 810–818 (2008).
10. Hatzenpichler, R. *et al. Proc. Natl Acad. Sci. USA* **105,** 2134–2139 (2008).
11. Ouverney, C. C. & Fuhrman, J. A. *Appl. Environ. Microbiol.* **66,** 4829–4833 (2000).
12. Herndl, G. J. *et al. Appl. Environ. Microbiol.* **71,** 2303–2309 (2005).
13. Kirchman, D. L. *et al. Limnol. Oceanogr.* **52,** 495–507 (2007).
14. Ingalls, A. E. *et al. Proc. Natl Acad. Sci. USA* **103,** 6442–6447 (2006).

ASTROPHYSICS

Quest for a habitable world

Drake Deming

A neat technique, applied to the brightest transiting extrasolar planet known, identifies an indisputable signature of water vapour in the planet's atmosphere. The method might be used to probe a nearby habitable world.

When the first discovery[1] of a planet orbiting a Sun-type star was announced in 1995, few could have imagined that we would stand where we do today: more than 300 planets outside our Solar System (exoplanets) are now known. The presence of most of these exoplanets has been inferred either through detecting their parent star's 'wobble', which is caused by the gravitational tug of the planet, or through measuring a dip in the host star's light as the planet passes in front of it (transits). The focus of exoplanet research is now shifting towards finding planets that are the most promising candidates for supporting life, and searching for indicators of biological processes (biomarkers) in their visible and infrared spectra. In this issue (page 767), Grillmair *et al.*[2] report that, by using long-duration observations from NASA's Spitzer Space Telescope, they have found an unequivocal signature of water vapour in the atmosphere of the brightest transiting exoplanet yet detected, the gas giant HD 189733b.

The transit method of detecting exoplanets has been especially fruitful for studying these objects because it allows the planet's light to be measured using a geometric trick. Almost all planets that transit their stars will also pass behind the star at a known time. Measuring the system before or after this 'secondary' eclipse gives the light of the star plus its planet, but a measurement of the system when the planet is behind the star gives the star's light in isolation. Subtracting one measurement from the other reveals the contribution of the planet.

The secondary-eclipse method was the technique used to measure the first crude spectra of HD 189733b and another extrasolar planet, based on Spitzer observations[3,4]. But these pioneering measurements raised more questions than they answered. We know that oxygen is the most abundant element in the cosmos after hydrogen and helium, and that oxygen atoms should combine with hydrogen in gas-giant planets to produce water vapour. But the first spectra of HD 189733b did not show the expected signature of water vapour. Was the water missing, variable in abundance, or just hidden by high cloud layers? The reason for its absence in the first observations is still unclear. Other investigations have found signs of water in giant exoplanets[5–7], but doubts have been slow to dissipate.

Grillmair *et al.*[2] blast away any remaining uncertainty by pushing the secondary-eclipse technique to its limit — averaging the results from ten eclipses of HD 189733b. The fruit of their persistence is a characteristic peak in the planet's spectrum that arises from a specific bending mode of the water molecule. The authors show conclusively that water vapour is present in the atmosphere of this hot gas giant world, validating a key aspect of our current models for these kinds of planets.

Giant hot exoplanets are fascinating, but what we ultimately want to find are life-bearing worlds. That search should target small, rocky or icy exoplanets, not gas giants, because the thick hydrogen–helium envelopes of gas giants are not where we expect to find life. The detection of atmospheric biomarkers will be crucial to finding life. However, although water is deemed to be necessary for life, atmospheric water vapour is not a biomarker; it can be seen in a wide variety of astrophysical environments, even in sunspots[8]. Nevertheless, we must first learn how to detect abundant molecules such as water before we can advance to identifying the more subtle signatures that scarcer molecules such as molecular oxygen leave in exoplanet spectra. Grillmair and colleagues have taken that first step.

But how do we proceed to find biomarkers in the atmosphere of a habitable exoplanet? The most obvious way would be to image a planet that is a twin of Earth — that is, one with an Earth-like mass and distance from its parent star — orbiting a nearby Sun-type star. This could be achieved by using high-contrast imaging observations to separate the planet's light from the bright glare of its star. Astronomers have successfully imaged massive planets orbiting far from their stars[9,10], but imaging a twin of our Earth will be a much more difficult task because terrestrial planets are orders of magnitude fainter and closer to the glare of their star. Using imaging would also allow us to obtain spectra and thereby find a potentially habitable world, but such a find seems many years away.

Fortunately, there is a quicker way to identify potentially life-bearing worlds in our close cosmic neighbourhood. Our nearest stellar neighbours are predominantly red-dwarf stars. These small, cool stars seem to lack gas-giant planets, but harbour rocky 'super-Earth' planets in abundance. Moreover, the habitable zone — the area around a star where water is in liquid form and life is potentially sustainable — is tucked in close to these stars. This is exactly the region where, as viewed from Earth, planets are more likely to transit the star. The odds seem good that in the next few years we will find a super-Earth exoplanet that transits a nearby red-dwarf star in, or close to, its star's habitable zone.

The search for a transiting super-Earth has already begun. David Charbonneau and his collaborators recently inaugurated the MEarth (pronounced 'mirth') Project[11]. MEarth uses a collection of 0.4-metre-diameter robotic ground-based telescopes to monitor the nearest 2,000 red dwarfs visible from the Northern Hemisphere in the search for the transit of a rocky planet. MEarth is fully operational at

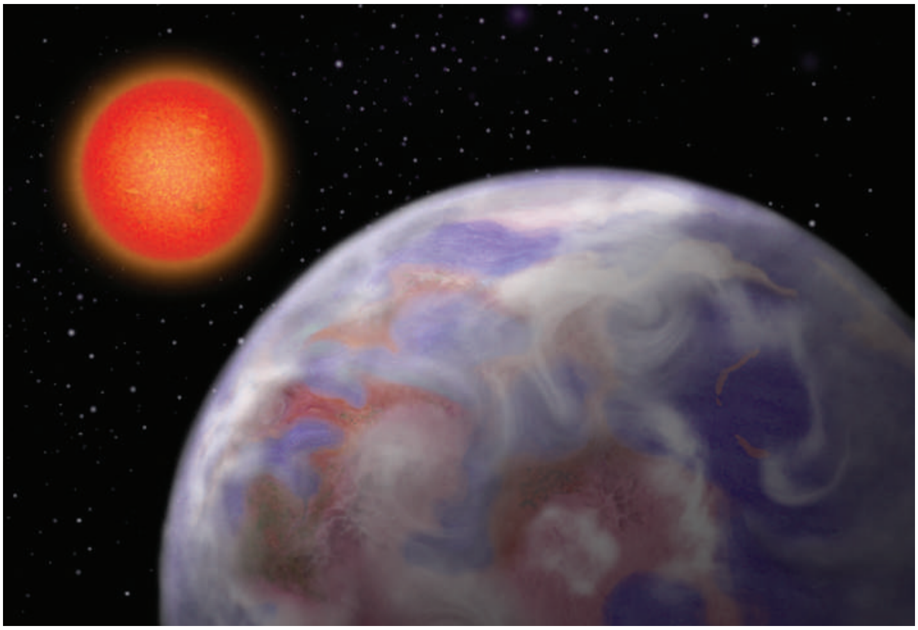


Figure 1 | A home for life as we don't know it? A rocky super-Earth transiting a nearby red-dwarf star may be the best short-term chance of finding a life-bearing world outside our Solar System. Grillmair and colleagues' technique[2] to measure spectra of exoplanets, which they applied to gas giant HD 189733b, the brightest transiting extrasolar planet yet detected, could ease that task. (Artwork by N. Spencer.)

the Fred Lawrence Whipple Observatory in Arizona, and the robotic telescopes survey our nearest stellar neighbours every clear night.

Also, NASA's prospective Transiting Exoplanet Survey Satellite mission could potentially find a rocky habitable planet transiting a nearby small star. Once we find such a planet, we will turn to the forthcoming James Webb Space Telescope to measure its spectrum, using the technique that Grillmair *et al.*[2] have pioneered. Orbiting a red dwarf, this habitable super-Earth may be a bizarre and unfamiliar world (Fig. 1), hosting 'life as we *don't* know it'. Many investigators think that this prospect is as exciting as finding a twin of our Earth, whose rarity is still an open question. ■

Drake Deming is in the Solar System Exploration Division, NASA Goddard Space Flight Center, Code 693, Greenbelt, Maryland 20771, USA.
e-mail: leo.d.deming@nasa.gov

1. Mayor, M. & Queloz, D. *Nature* **378,** 355–359 (1995).
2. Grillmair, C. J. *et al. Nature* **456,** 767–769 (2008).
3. Grillmair, C. J. *et al. Astrophys. J.* **658,** L115–L118 (2007).
4. Richardson, L. J. *et al. Nature* **445,** 892–895 (2007).
5. Tinetti, G. *et al. Nature* **448,** 169–171 (2007).
6. Swain, M. R., Vasisht, G. & Tinetti, G. *Nature* **452,** 329–331 (2008).
7. Barman, T. S. *Astrophys. J.* **676,** L61–L64 (2008).
8. Wallace, L. *et al. Science* **268,** 1155–1158 (1995).
9. Kalas, P. *et al. Science* **322,** 1345–1348 (2008).
10. Marois, C. *et al. Science* **322,** 1348–1352 (2008).
11. Irwin, J., Charbonneau, D., Nutzman, P. & Falco, E. in *Proc. IAU Symp.* **253** (in the press).

NEUROSCIENCE

Brain power

Catherine N. Hall and David Attwell

The brain relies on blood to supply oxygen and glucose for energy. Surprisingly, neuronal activity, acting via supporting cells called astrocytes, can increase or decrease blood flow depending on whether oxygen levels are low or high.

The air entering our lungs contains 21% oxygen, and by the time the oxygen reaches our brain this value has fallen to 4%. Yet physiological experiments on isolated brain tissue traditionally use solutions containing 95% oxygen. On page 745 of this issue, Gordon *et al.*[1] demonstrate how this experimental oversupply of oxygen may mislead us about the mechanisms that are physiologically important in the brain, and at the same time they provide a notable advance in our understanding of how the brain regulates its energy supply.

When an area of the brain is active, it needs more energy to power information processing. Information flow along and between neurons is based on the movement of ions across the cell membrane, and, to generate the energy required for pumping ions out, neurons depend largely on glucose and oxygen supplied by the blood.

One way in which neurons control their energy supply is by molecular signalling to surrounding support cells called astrocytes. This signalling is mediated by the neurotransmitter glutamate, which raises the calcium ion (Ca^{2+}) concentration in astrocytes, thereby generating arachidonic acid from membrane lipids. Surprisingly, arachidonic acid can either dilate[2,3] or constrict[4] the blood arterioles to increase or decrease blood flow, respectively. Dilation follows the conversion of arachidonic acid to prostaglandin E_2 and epoxygenase derivatives, which diffuse to the smooth-muscle cells encircling nearby arterioles and relax them. Constriction, by contrast, is caused by arachidonic acid itself diffusing to the smooth muscle, where its derivative 20-HETE causes arteriole constriction by increasing Ca^{2+} entry through voltage-gated channels[4].

What determines whether arterioles dilate or constrict in response to a high Ca^{2+} concentration in astrocytes? Gordon *et al.*[1] find that the prevailing oxygen level is a key determinant of whether astrocytes increase or decrease blood flow.

Two mechanisms seem to mediate this oxygen-dependent switch. First, at low oxygen levels, increased energy generation in neurons and astrocytes by a pathway that does not use oxygen (glycolysis) results in the extracellular accumulation of lactate. Because transporter proteins in the cell membrane carry prostaglandins into cells in exchange for lactate[5], lactate levels affect how prostaglandins are removed from the extracellular space. A high extracellular concentration of lactate inhibits prostaglandin removal, enhancing extracellular prostaglandin levels and so dilating arterioles[1].

Second, because at low oxygen levels there is less energy available to make the high-energy compound ATP, intracellular concentrations of the ATP precursor adenosine rise[1]. Adenosine is then transported out of neurons and astrocytes, and inhibits Ca^{2+} entry into the arteriole smooth muscle through voltage-gated channels, preventing the constriction that its entry would induce. Thus, low oxygen levels promote increased blood flow by simultaneously potentiating the dilation pathway and inhibiting the constriction pathway in astrocytes[1].

As with all major advances, Gordon and colleagues' study raises as many questions as it answers. For example, how much lactate is needed to promote dilation? The authors' data[1] suggest that physiological levels of extracellular lactate (about 1 millimolar) in the brain can inhibit prostaglandin uptake, yet when prostaglandin transporters are expressed in a cancer cell line[5], much higher lactate concentrations (48 mM) are needed to produce 50% inhibition. Conceivably, the transporter's affinity for lactate is higher in astrocytes than in the cell line. It is noteworthy that lactate can also dilate arterioles

by a prostaglandin-independent mechanism[6].

Gordon *et al*. demonstrate astrocyte-mediated dilation of arterioles in solutions containing 20% oxygen, which produces approximately physiological levels of oxygen in brain slices[7]. They also find that astrocyte-mediated constriction occurs in solutions containing oxygen levels well above the physiological (95%). Why has evolution produced the latter pathway, in which neural activity decreases blood flow? And will physiological tissue concentrations of oxygen ever be high enough to activate this pathway? It turns out that 20-HETE-mediated arteriole constriction is inhibited[8] by nitric oxide (NO), a molecule that is released by neurons in response to glutamate secreted by neighbouring neurons (and which can also directly dilate arterioles). The 20-HETE-mediated pathway may therefore be better viewed as a mechanism producing a basal constriction of arterioles that can then be modulated by NO to provide another pathway for activity-dependent dilation.

Future work is likely to focus on how changes in the levels of lactate, adenosine, oxygen and NO interact to coordinate blood flow and hence the brain's energy supply. Some clues can be found in previous data. For example, NO released by neurons inhibits the conversion of arachidonic acid to epoxygenase derivatives that evoke dilation[9]. As NO production in neurons requires oxygen[10], at low oxygen levels this mechanism will be inhibited, promoting dilation. Moreover, oxygen is needed for the synthesis of both constricting (20-HETE) and dilating (prostaglandin E_2 and epoxygenase) derivatives of arachidonic acid. At low oxygen levels, however, the production of 20-HETE is inhibited more strongly than that of prostaglandin E_2 and epoxygenase derivatives[8], increasing dilation. Finally, in blood capillaries, where contractile cells called pericytes may regulate blood flow[11], lactate causes constriction at high oxygen levels, but dilation at low levels[12]. There is, therefore, an array of switching mechanisms that promote brain energy supply when oxygen levels fall.

In a wider context, Gordon and colleagues' observations raise questions for both cognitive neuroscientists and neurologists. Could the initial dip in local oxygen concentration that accompanies neural activity[13] affect astrocyte signalling rapidly enough to contribute to the increase in blood flow that generates the signals seen in functional imaging of the brain? And could our new understanding of astrocyte signalling lead to better therapies for correcting disorders of blood flow in the brain, such as those that occur after stroke and in vascular dementia?

Gordon *et al.*[1] have opened a fresh chapter in our investigation of how blood flow is regulated in the brain. But their work has a broader implication: physiological studies using solutions bubbled with 95% oxygen may be altering the operation of signalling pathways in the brain, producing misleading results. ■

Catherine N. Hall and David Attwell are in the Department of Neuroscience, Physiology and Pharmacology, University College London, London WC1E 6BT, UK.
e-mail: d.attwell@ucl.ac.uk

1. Gordon, G. R. J., Choi, H. B., Rungta, R. L., Ellis-Davies, G. C. R. & MacVicar, B. A. *Nature* **456,** 745–749 (2008).
2. Zonta, M. *et al. Nature Neurosci.* **6,** 43–50 (2003).
3. Takano, Y. *et al. Nature Neurosci.* **9,** 260–267 (2006).
4. Mulligan, S. J. & MacVicar, B. A. *Nature* **431,** 195–199 (2004).
5. Chan, B. S., Endo, S., Kanai, N. & Schuster, V. L. *Am. J. Physiol. Renal Physiol.* **282,** F1097–F1102 (2002).
6. Hein, T. W., Xu, W. & Kuo, L. *Invest. Ophthalmol. Vis. Sci.* **47,** 693–699 (2006).
7. Hall, C. N. & Attwell, D. *J. Physiol. (Lond.)* **586,** 3597–3615 (2008).
8. Roman, R. *J. Physiol. Rev.* **82,** 131–185 (2002).
9. Metea, M. R. & Newman, E. A. *J. Neurosci.* **26,** 2862–2870 (2006).
10. Stuehr, D. J., Santolini, J., Wang, Z., Wei, C. & Adak, S. *J. Biol. Chem.* **279,** 36167–36170 (2004).
11. Peppiatt, C. M., Howarth, C., Mobbs, P. & Attwell, D. *Nature* **443,** 700–704 (2006).
12. Yamanishi, S., Katsumura, K., Kobayashi, T. & Puro, D. G. *Am. J. Physiol. Heart Circ. Physiol.* **290,** H925–H934 (2006).
13. Vanzetta, I. & Grinvald, A. *NeuroImage* **13,** 959–967 (2001).

MATERIALS SCIENCE

Deformation of the ultra-strong

Subra Suresh and Ju Li

In situ electron microscopy observations of the extrusion of single nanocrystals from graphitic cages show that these crystals deform near their theoretical strength limits. The question is how this happens.

Many nanostructured materials can sustain specimen-wide stress up to more than a tenth of their ideal strength for a considerable time. For example, a test performed on a monolayer of graphene yielded a strength value very close to its ideal strength calculated by quantum mechanics[1]. But the way that these ultra-strong materials respond to deformation at high temperatures remains mysterious because performing temperature-controlled mechanical tests of nanostructures *in situ* is not easy.

Writing in *Physical Review Letters*[2], Sun *et al.* report observations of plasticity — the permanent, irreversible deformation of a material, as opposed to elastic deformation in which atomic bonds are stretched but not broken — of nanometre-sized metallic crystals inside a transmission electron microscope (TEM). They ascribe the observed deformation to the activity of short-lived, string-like defects in the crystals, known as dislocations.

In 1926, Jacov Frenkel estimated[3] the ideal (maximum attainable) shear strength of a perfect crystal to be about a tenth of its shear modulus (initial rigidity). But at that time, tests performed on real materials yielded strengths two to three orders of magnitude lower. This discrepancy was attributed to dislocations, which are boundaries of planar fault regions in the crystal structure where atoms slip out of position when the material is strained. But dislocations were directly observed in the TEM only 30 years later[4]. Once created, dislocations move and multiply easily on their own as the material is subjected to loading. Common metal objects — for example, a kitchen fork — contain many dislocations to start with, and thus deform at stresses much lower than their ideal shear strengths.

Relatively large characteristic structural dimensions, such as micrometre-sized grains in bulk materials, facilitate the continuous generation, entanglement and storage of dislocations during plastic deformation. However, as the characteristic scale (such as the crystal grain size or the smallest dimension of a thin film) shrinks below 100 nm, dislocations are 'fatally attracted' to internal interfaces (such as crystal grain boundaries) and surfaces of the specimen. Consequently, it becomes much more difficult to sustain a permanent population of mobile dislocations — which are the vehicles of plastic deformation during straining — inside the material[5,6]. In these cases, deformation can be achieved only if new dislocations are nucleated afresh, usually from the same internal interfaces and surfaces that also absorb and annihilate them[7,8]. The continual need to nucleate new dislocations in these tiny crystals results in a significant increase in the material's strength.

In their experiment, Sun and colleagues[2] initially confined individual, three-dimensional crystals (as small as 10 nm in diameter) of materials, including gold and platinum, in spherical graphitic shells. Subsequent puncturing and irradiation of the shells by a focused electron beam at different temperatures led to the extrusion of the crystals from the capsules (Fig. 1a, b). From direct comparison of the lattice spacings in the gold nanocrystals inside and outside the capsules, the authors inferred a prevailing pressure of about 20 gigapascals (about 200,000 times the standard atmospheric pressure) in the capsule when this was irradiated at about 300 °C. This is an extremely high stress for gold, considering that its ideal shear strength is only about 1 gigapascal. There is thus no question that these systems are ultra-strong.

More controversial, however, is the mechanism of deformation during extrusion. With

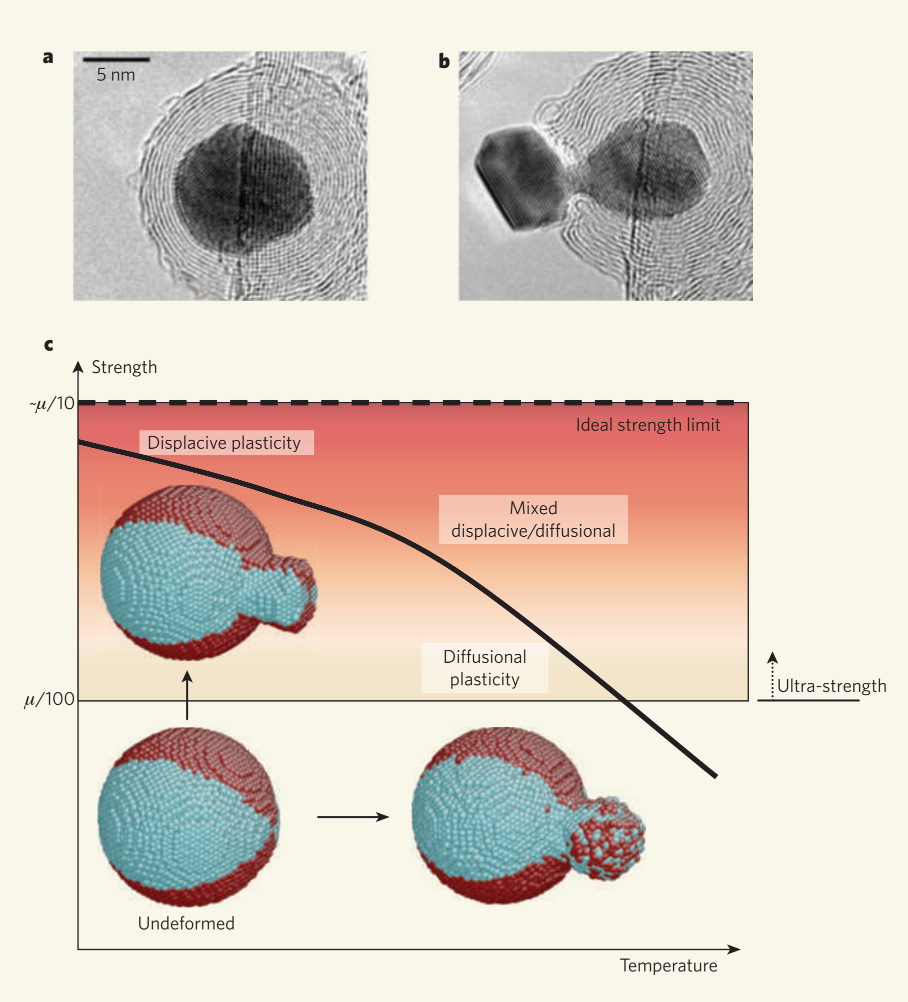


Figure 1 | Crystal plasticity. Sun *et al.*[2] performed *in situ* transmission electron microscopy observations of the extrusion of single gold nanocrystals from graphitic capsules under electron irradiation at 300 °C. **a**, Before irradiation. **b**, After irradiation for 540 seconds. **c**, Deformation mechanisms. The black curve shows the typical dependence of the strengths of crystalline materials — expressed as a fraction of their shear modulus, μ — on temperature. As the temperature increases, one of three competing mechanisms operates: displacive, mixed or diffusional plasticity. Superimposed are illustrative simulations, which we carried out, of the plasticity of copper nanospheres at a temperature of 300 K (sphere at the top) and 900 K (sphere on the right), in which deformation is thought to be controlled by displacive and diffusional plasticity, respectively. The copper atoms are shown in two colours (red and cyan) to make it easier to track their motions from the undeformed crystal state (bottom left sphere) to the deformed state. At 900 K, the random mixing of red- and cyan-coloured atoms in the extrusion-neck region (where the stress gradient is largest) indicates that extensive surface diffusion is taking place. (**a**, **b**, Courtesy American Physical Society.)

a TEM, Sun *et al.* observed a highly perfect atomic structure with occasional grain boundaries and planar stacking faults. But dislocations were not visible. On the basis of molecular-dynamics simulations, Sun *et al.* conclude that deformation originates from individual, transient dislocations that are freshly nucleated and vanish so fast that they cannot be seen with a TEM. Although diffusive atomic processes could be active at 300 °C in gold, the authors argue that diffusion does not contribute to plastic strain, and that the observed strength and deformation can be accounted for solely by the nucleation and motion of short-lived dislocations.

One of three competing mechanisms, all dependent on temperature and mechanical strain rate, induces plastic deformation: displacive, diffusional or mixed plasticity. Displacive plasticity[5,7,8] is produced by the collective shearing of atoms, that is, the glide of dislocations. Diffusional plasticity[9] is governed by many, almost random, individual atom or vacancy hops. In conventional coarse-grained metals, typically below about $T_M/3$, where T_M is the absolute temperature at melting, deformation is dominated by displacive mechanisms, whereas above about $2T_M/3$ diffusional mechanisms control the process. A mixture of these two mechanisms occurs at in-between temperatures; in such cases the inelastic strain is still mainly produced by dislocation glide but its rate is controlled by diffusion (Fig. 1c).

Lack of understanding of the deformation mechanisms that can operate in ultra-strong materials severely limits our ability to create nanometre-scale systems with the desired mechanical properties. Information about deformation mechanisms is often gathered from molecular-dynamics simulations, but these are limited to unrealistically high strain rates. Recently, progress has been made through the use of computational methods that elucidate mechanisms of displacive plasticity at low temperatures through direct calculations of the activation volume, which characterizes the sensitivity of plastic-yield stress (the stress at which the material deforms permanently) to strain rate. Such computational studies reveal that low-temperature deformation of ultra-strong systems, such as the nanocrystals studied by Sun *et al.*, become highly sensitive to strain rate and temperature[7,8]. The underlying mechanism involves the nucleation, absorption and desorption of dislocations from interfaces and free surfaces, with a resultant reduction in activation volume, typically 2–20 times the volume of a single atom (Ω_0). This activation volume is much smaller than those observed for traditional displacive-plasticity mechanisms (about $10^3\,\Omega_0$) that operate in coarse-grained polycrystals. It is, however, still larger than those of typical vacancy processes, for which the activation volume is less than about Ω_0.

But at higher temperatures, such as 300 °C in gold, the way deformation changes with strain rate and the scale of nanostructures is unknown. In particular, the temperature and stress boundaries that separate the displacive processes from the diffusional and mixed processes will shift from those of the corresponding coarse-grained materials. Further experiments and modelling at higher temperatures[9–11] will inevitably be needed to understand deformation in nanostructured materials. Meanwhile, Sun *et al.*[2] have developed an innovative *in situ* experimental method that could provide insight into the process. ■

Subra Suresh is in the School of Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
Ju Li is in the Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mails: ssuresh@mit.edu; liju@seas.upenn.edu

1. Lee, C., Wei, X., Kysar, J. W. & Hone, J. *Science* **321,** 385–388 (2008).
2. Sun, L., Krasheninnikov, A. V., Ahlgren, T., Nordlund, K. & Banhart, F. *Phys. Rev. Lett.* **101,** 156101 (2008).
3. Frenkel, J. *Z. Phys.* **37,** 572–609 (1926).
4. Hirsch, P. B., Horne, R. W. & Whelan, M. J. *Phil. Mag.* **1,** 677–684 (1956).
5. Greer, J. R. & Nix, W. D. *Phys. Rev. B* **73,** 245410 (2006).
6. Shan, Z. W., Mishra, R. K., Syed Asif, S. A., Warren, O. L. & Minor, A. M. *Nature Mater.* **7,** 115–119 (2008).
7. Zhu, T., Li, J., Samanta, A., Kim, H. G. & Suresh, S. *Proc. Natl Acad. Sci. USA* **104,** 3031–3036 (2007).
8. Zhu, T., Li, J., Samanta, A., Leach, A. & Gall, K. *Phys. Rev. Lett.* **100,** 025502 (2008).
9. Wolf, D., Yamakov, V., Phillpot, S. R., Mukherjee, A. & Gleiter, H. *Acta Mater.* **53,** 1–40 (2005).
10. Mason, J. K., Lund, A. C. & Schuh, C. A. *Phys. Rev. B* **73,** 054102 (2006).
11. Huang, J. Y. *et al. Nature* **439,** 281 (2006).

OBITUARY

Martin Lindauer (1918–2008)

Prime mover in behavioural physiology and sociobiology.

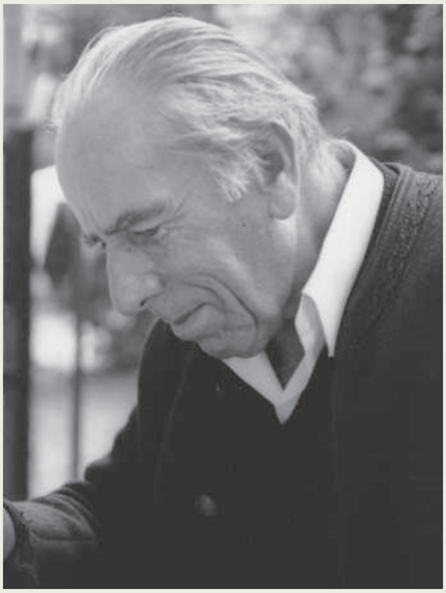
Martin Lindauer, who died on 13 November 2008, was a leading light in the discovery of how honeybees communicate and learn, sense the world, find their way, and live in societies. His enduring influence will also be felt through his success in nurturing a generation of German neurobiologists and behaviourists during his tenure, from 1963 to 1973, as a professor of zoology at the University of Frankfurt.

Lindauer was born in a tiny village in the foothills of the Bavarian Alps, the next to youngest of 15 children in a poor farming family. Here he grew up close to nature — including the bees in his father's hives — but he was also an outstanding student, and won a scholarship to a distinguished Bavarian boarding school in Landshut. In April 1939, he was drafted into Hitler's Work Service to dig trenches, but six months later was transferred to the army and assigned to an anti-tank unit. In July 1942, on the Russian front, he was badly wounded. This was to prove his salvation. He was removed from the front, but the other 156 men in his company went on to the Battle of Stalingrad. Only three returned alive.

While recovering in Munich, Lindauer attended a lecture delivered by Karl von Frisch, a distinguished zoology professor at the university. Lindauer later recalled the occasion, saying that when he heard von Frisch talking about cell division, he felt he had returned to "a world of humanity", where humans create rather than destroy. He resolved to study biology, ultimately starting his PhD research on honeybees, with von Frisch as his adviser, in the spring of 1945.

The previous summer, von Frisch had made a revolutionary discovery, one for which he would eventually receive the Nobel prize: an insect, the worker honeybee, can inform her hive mates of the direction and distance to a rich food source by means of dance behaviour. Von Frisch wanted to press forward with this research, so he assigned Lindauer a relevant (but less exciting) topic, namely, how these communication dances are affected by food scent and taste.

Lindauer had a knack for noticing some anomaly or behavioural quirk that would turn out to be important. So, while pursuing his assigned topic on how bees tune their dancing in relation to the properties of a food source (nectar sweetness, abundance and so on), he noticed that the bees' assessment of a food source's desirability, as indicated by their willingness to advertise the source, depends not only on its profitability as an energy source, but also on how badly the colony


H. HEILMANN

needs the food the bee is providing.

On cool, cloudy days, when the delivery of nectar to the hive was slow, Lindauer's bees would dance for the slightest taste of sugar; by contrast, on hot, sunny days, when nectar was flowing heavily into the hive, his bees would dance only if he made his feeder super sweet. How could a forager, working out in the fields, keep herself informed of her colony's changing needs? Lindauer discovered the answer: the speed with which a forager is unloaded of nectar tells her whether the colony's need is high (fast unloading) or low (slow unloading). These findings proved a springboard for numerous studies of how 'queuing delays' are used to control processes in animal societies.

Between 1945 and 1960, Lindauer repeatedly demonstrated his ability to detect the subtle features that enable coherence to emerge from the apparent tumult in a beehive. He discovered how individuals stay informed of their colony's labour needs. This process rarely involves bee-to-bee signalling; instead, each bee is mostly her own informant, devoting much time to reconnoitring inside the hive. He discovered how a colony regulates its water intake to avoid dehydration: when the need for water increases, any bee that enters the hive with a droplet of water receives a stormy welcome, with several hive bees eagerly taking her water from her, and she is thereby inspired to dance and activate additional water collectors. And he discovered how a honeybee swarm chooses its new home: house-hunting bees fly out, locate several candidate sites, then return to the swarm where they perform dances to advertise their findings to uncommitted house hunters. The more highly a site is valued, the more strongly it is advertised and the more supporters it gains. Eventually, the best site dominates the discussion and is occupied.

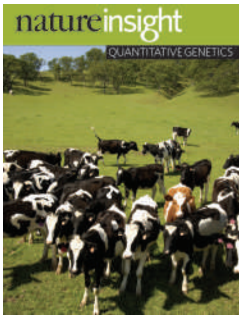
While blazing trails in sociobiology in his own studies, Lindauer was also solving mysteries in behavioural physiology in collaboration with von Frisch. Their most notable joint project was investigating whether bees can calculate, using knowledge of the time of day and the Sun's diurnal course, a specific compass direction according to the Sun's position. Von Frisch doubted that bees were capable of this. But Lindauer persisted, and devised a clever experiment for testing the idea: bees were trained to find food in one direction (south, say) on one afternoon, then were taken overnight to an unfamiliar location and tested the next morning for the ability to orient to the south to find food. Lindauer found that they did indeed fly south on the morning of the test.

In 1963, Lindauer accepted a professorship at the University of Frankfurt, where he attracted a constellation of students — including Bert Hölldobler, Eduard Linsenmaier, Hubert Markl, Randolf Menzel and Rüdiger Wehner — who became his successors in advancing behavioural physiology and sociobiology in Germany. Lindauer nurtured his students by giving them the freedom to prove their talent, providing due praise and encouragement, and also guidance in writing clear and enjoyable scientific papers. Lindauer left Frankfurt in 1973 to take up a professorship at the University of Würzburg. There he devoted himself primarily to administrative work until his retirement in 1987.

In summarizing his work, Lindauer published a small volume entitled *Communication Among Social Bees*, based on lectures he had delivered at Harvard University. In the preface, he expressed his pleasure in discovering how "common problems and interests within biology [had] built new bridges ... that span the wrongs of the past". His life's work leaves us not just with a legacy of discoveries, but also with the lesson that the study of nature, such as the social order of bees, can bring sweetness out of chaos.

Thomas D. Seeley
Thomas D. Seeley is in the Department of Neurobiology and Behavior, Cornell University, Ithaca, New York 14853, USA.
e-mail: tds5@cornell.edu

QUANTITATIVE GENETICS



Cover illustration
Animal coat colour is a complex trait resulting from small contributions from many genes. (Image courtesy of P. Schermeister/ Photolibrary.com. Artwork by N. Spencer)

Editor, *Nature*
Philip Campbell
Publishing
Nick Campbell
Samia Burridge
Claudia Banks
Insights Editor
Lesley Anson
Production Editor
Davina Dadley-Moore
Senior Art Editor
Martin Harrison
Art Editor
Nik Spencer
Sponsorship
Amélie Pequignot
Production
Jocelyn Hilton
Marketing
Elena Woodstock
Emily Elkins
Editorial Assistant
Alison McGill

The father of modern quantitative genetics, R. A. Fisher, is credited with the observation: "Natural selection is a mechanism for generating an exceedingly high degree of improbability". The truth in this is evident just from looking at evolutionary history — the improbable happens all the time. It is made possible by large reservoirs of genetic variation, the occurrence of mutations, and interactions between genes and the environment. Recent revolutions in genomics have allowed genetic variation to be described at an unprecedented level, pinpointing nucleotide changes that directly or indirectly affect phenotype. The result is a renaissance in quantitative genetics, particularly in the study of complex traits, enabling theories to be tested by experiment on a genome-wide scale.

One now common approach for examining complex traits in humans is the genome-wide association study, which compares variation across the genomes of two groups of individuals and looks for markers that are associated more with one group than the other, a strategy now being applied to plants and other organisms. After the variants linked to a trait have been identified by such approaches, the effects of these variants on the molecular networks (such as gene expression) in a cell or organism can be investigated. This Insight highlights progress in teasing apart the basis of complex traits in this way, as well as in determining the relative contributions of genetics and the environment to traits in various organisms, from plants to mice to the trio of the malaria parasite, its mosquito vector and the human host.

Genetics has always been a field in which studies in one organism are crucial for discoveries in another, and we hope that the articles in this Insight inspire crosstalk between specialists, as well as entice scientists from other fields. We are pleased to acknowledge the financial support of Monsanto, which contributed towards the distribution of this Insight. As always, *Nature* carries sole responsibility for editorial content and peer review.

Chris Gunter, Consultant Editor

PROGRESS

COMMENTARIES

REVIEW

nature insight

NATURE|Vol 456|11 December|doi:10.1038/nature07629

Next-generation genetics in plants

Magnus Nordborg[1] & Detlef Weigel[2]

Natural variation presents one of the fundamental challenges of modern biology. Soon, the genome sequences of thousands of individuals will be known for each of several species. But how does the genotypic variation that will be observed among these individuals translate into phenotypic variation? Plants are in many ways ideal for addressing this question, and resources that are unmatched, except in humans, have now been developed.

When it comes to dissecting complex traits, genetic studies of plants have always been at the forefront. Work with cereals provided the first demonstrations that segregation at multiple loci could give rise to a continuous distribution of phenotypes[1,2], and it was in peas that individual quantitative trait locus (QTL) effects were first inferred using markers[3]. Many genes responsible for QTL effects have now been cloned in plants, mainly as a result of the persistent application of classical genetic techniques. And it is not just the 'flagship' of plant biology, *Arabidopsis thaliana*, that has proved useful in such studies. Some of the earliest successes came from studying crop plants, such as maize (corn), rice and tomato[4–7].

The genetic maps of many organisms are now becoming increasingly dense, and the cost of genotyping is decreasing. For these reasons, it has become almost routine to identify the genes (or QTLs), and even the individual nucleotides (QTNs), that cause particular phenotypic effects, using a process that starts with linkage mapping (conventional genetic mapping based on the idea that the farther apart two linked genes are, the more likely a recombination event between them will be) in populations derived from crosses between divergent strains. Notably, despite the extensive genetic screens for mutants of *A. thaliana* affecting a wide variety of phenotypes, the above approach has enabled plant geneticists to identify numerous QTL genes, which were not previously known to have an impact on the trait examined (see refs 8–11 for examples).

High-resolution linkage mapping is slow and labour intensive, however, and for all the success stories, there are probably at least as many cases in which QTL cloning efforts were abandoned (and therefore not reported in the literature) because of difficulty with fine mapping, complex genetic architecture and so on. Therefore, there is great interest in developing an alternative technique — genome-wide association (GWA) mapping — which looks for associations between phenotypes of interest and the DNA sequence variants present in an individual's genome, as assessed by determining an individual's genotype at the positions of hundreds of thousands of single nucleotide polymorphisms (SNPs). GWA studies provide much higher resolution than linkage mapping because they involve studying a natural population rather than the offspring of crosses, and associations in natural populations are typically on a much finer scale because they reflect historical recombination events. Here we describe the recent progress in realizing the potential of GWA mapping in plants, and we then discuss the rapid advances that are expected during the next year or two.

Considerations for GWA mapping

Many plants are well suited for GWA studies, in particular species that self-fertilize, such as *A. thaliana* and rice (*Oryza sativa*), and species that can be clonally propagated, such as switchgrass (*Panicum virgatum*) and grape (*Vitis vinifera*). This is because after lines of these species have been densely genotyped or completely sequenced, the plant genetics community can analyse an unlimited number of traits in genetically identical material. This ability to examine individuals with identical genotypes greatly improves statistical power when studying phenotypes with complex genetics, especially when the phenotypes are modified by the environment.

With association mapping, however, the structure of the population can be a strong confounding factor (discussed later), especially for traits that are important in adaptation to the local environment. There are statistical solutions to this problem, but it is important to recognize that these do not always work and that linkage mapping in controlled crosses (which is robust to confounding) is sometimes necessary, perhaps as a complement to GWA mapping (Fig. 1).

Another consideration is that these two mapping techniques — association mapping and linkage mapping — also differ in terms of how the genetic architecture of the trait affects statistical power (Fig. 2). The importance of a particular allele in phenotypic variation across a population depends on its frequency, as well as on its effect. Thus, GWA mapping has low power to detect rare alleles, even if these alleles have a large phenotypic effect (Fig. 2c). Conversely, alleles that are identified by mapping in crosses between two essentially randomly chosen parents can have a large effect but might not be important from an evolutionary perspective because they are rare. In other words, mapping QTLs by using crosses might bias researchers towards identifying rare (and often perhaps deleterious) alleles that have large effects but little relevance to most of the phenotypic diversity found in nature.

Resources for GWA mapping

The development of reference resources for genomic polymorphism in *A. thaliana* mirrored the progress of the International HapMap Project, which identified SNPs in the human population worldwide[12,13], and resources are now at a stage similar to those for humans. To build these resources, a pilot study was carried out to investigate the basic pattern of polymorphism in *A. thaliana* (that is, the level of variability, extent of linkage disequilibrium, and structure of the population): this involved direct sequencing of 1,400 loci in 96 individuals[14]. Then, a subset of 20 of these individuals was selected for resequencing using a high-density microarray[15]. The SNPs uncovered by this study[15] led to the design, in conjunction with Affymetrix, of a genotyping microarray with 250,000 probes for the purpose of GWA studies of *A. thaliana*[16]. This microarray is now being used to genotype about 1,300 naturally inbred lines (see Genomic polymorphism data in *Arabidopsis thaliana*, http://walnut.usc.edu/2010), which will be distributed to the community for extensive phenotyping. The project is likely to be a model for analogous projects in other species, including one that is underway to

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA. [2]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.

study rice (see the International Rice Functional Genomics Consortium, http://irfgc.irri.org).

As was the case for the International HapMap Project in humans, many approaches used for the study of *A. thaliana* and rice have already been rendered obsolete by technical advances. Future efforts to uncover SNPs will use next-generation sequencing approaches (such as Illumina's sequencing technology and Applied Biosystems's SOLiD System, which are already commercially available) rather than the microarray-hybridization technology that was used to construct the first-generation haplotype maps of *A. thaliana*[15] and rice, a technology that was more costly and less precise than newer approaches and was highly biased. In addition, rapidly decreasing costs mean that sophisticated schemes that have been used to select the most informative SNPs for genotyping are increasingly becoming irrelevant. By the time that researchers had determined that 140,000 tag SNPs (a subset of informative SNPs) would suffice to cover the *A. thaliana* genome, there was no economic reason not to genotype all 250,000 known high-quality SNPs that were not singletons (which are SNPs that have been detected only in single individuals and whose predictive power for other SNPs is therefore unknown)[16].

The importance of population structure

What, therefore, is the prospect of pinpointing individual genes with GWA approaches? It is well known that demography affects linkage disequilibrium. One example is that there is more linkage disequilibrium in Europeans than in Africans, reflecting humanity's African origins[12,13]. Another is that for wild *A. thaliana*, linkage disequilibrium is more extensive in North America than in Europe, consistent with the plant having been introduced into North America only after Europeans settled there[14,16]. In both cases, the probable explanation is that there was a bottleneck in colonization, with recombination not yet having had enough time to whittle down linkage disequilibrium among the alleles present on the limited number of founder chromosomes.

It is perhaps not as widely recognized that, in the presence of population structure, the genetic architecture of a trait in a sample of individuals depends on how the sample was assembled. For example, GWA mapping immediately reveals the importance of the gene *FRIGIDA* in the variation in flowering time among *A. thaliana* strains from the northwestern parts of continental Europe (where common loss-of-function alleles are an important determinant of early flowering) but not from central Asia (where no single loss-of-function allele is particularly frequent[17]). If variation in a trait is caused by numerous alleles of a single gene (as opposed to a small number of frequently occurring alleles), then researchers carrying out a GWA scan using global samples run the risk of concluding that there is no major locus for this trait (Fig. 2). This is essentially another facet of the problem with population structure that was mentioned earlier: the importance of a particular allele always depends on the reference population, and it is far from clear which population is meaningful from an evolutionary perspective.

Much attention has been given to population structure being a strong confounding factor in association studies, especially for traits that are important in local adaptation (such as flowering time in plants or skin colour in humans). Studies of maize and *A. thaliana*[18–22] have been at the forefront of identifying this problem and indicating statistical solutions. Application of one of these strategies[20] has already led to the identification of a major locus in maize that controls concentrations of provitamin A — an important trait, particularly for people with limited access to a diverse diet[23].

Combining association mapping and linkage mapping

A clear solution to the problem posed by population structure is to complement GWA studies of natural populations with linkage mapping of experimental populations, taking advantage of the increased resolution of the former and the robustness to confounding of the latter, a strategy that has been successfully applied to *A. thaliana*[11,22,24]. When studying human genetics, however, controlled crosses are not possible, so the solution is to use the transmission-disequilibrium test (TDT)[25], which uses the transmission of alleles from parents to offspring to verify linkage.

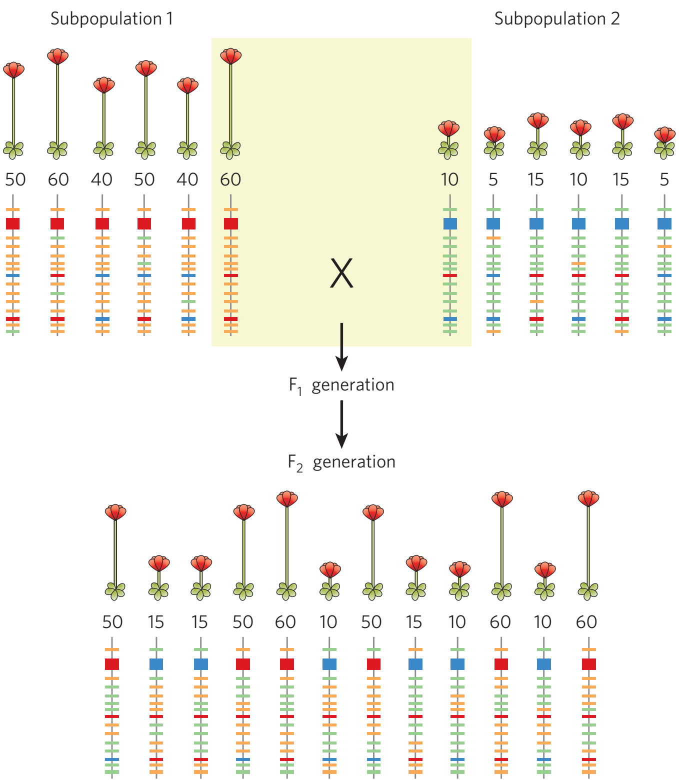


Figure 1 | GWA mapping is ineffective if there is strong genetic differentiation between subpopulations (that is, if there is structure in the population). In this example, two subpopulations of plants are depicted, one tall and one short (as illustrated and indicated by the numerical measurement), together with a schema of the genotype of each plant. The presence of red alleles increases the height of a plant, whereas blue alleles decrease the height; one locus has a major effect, and two have a minor effect. The many background markers (orange and green) are mostly exclusive to a specific subpopulation but are also strongly associated with height, even though they are not causal. By crossing the plants (shaded area) and generating an experimental population of F_2 generation or recombinant inbred lines, any linkage disequilibrium between background markers and causal markers is broken up, and the causal loci can then easily be mapped, albeit with relatively poor resolution.

For maize, by contrast, Ed Buckler and colleagues have pioneered a distinct approach, which is called nested association mapping[26]. GWA studies such as those underway in humans and *A. thaliana* would, at least for the next couple of years, be prohibitively expensive in maize, because its genome is larger than that of humans, is more polymorphic and has less-extensive linkage disequilibrium. Instead, 5,000 recombinant inbred lines (RILs) have been derived from separate crosses of a common standard genotype with 25 genetically diverse lines. The founder lines will be sequenced, whereas the RILs will be genotyped only with sufficient density to identify the ancestral founder at each point in the genome, resulting in a haplotype map that is essentially complete for each of the 5,000 RILs. Because crossing over during RIL formation is limited, such mapping can be accomplished with relatively high accuracy by using a moderate number of markers[26]. This mapping approach is conceptually similar to those applied to a heterogeneous stock of laboratory mice[27] or the Collaborative Cross[28] (a resource that is being generated with the aim of obtaining 1,000 RILs from eight standard mouse strains) (see page 724).

The nested-association-mapping design therefore, in effect, relies on the experimental crosses to map genes — without the confounding effects of population structure — to only a few, but still relatively large, genomic regions. Within these mapping intervals, allele sharing across the founder lines is exploited to achieve the resolution of GWA mapping. It is easy to see how this strategy could be applied to *A. thaliana* by the appropriate selection of subsets of lines from the many available RIL populations.

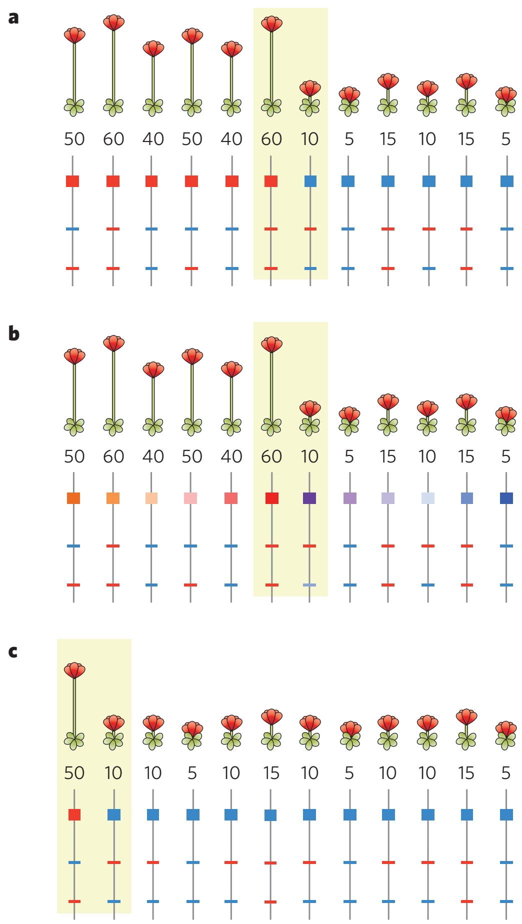


Figure 2 | The performance of GWA mapping depends strongly on the genetic architecture of the trait. Plants are depicted as in Fig. 1. Red alleles increase plant height, whereas blue alleles decrease height; one locus has a major effect, and two have a minor effect. **a**, Provided that there is no confounding effect of population structure, the major causal locus of a trait can easily be mapped by a GWA scan (otherwise an approach such as that in Fig. 1 is needed). **b**, Even if there is a major locus for the trait, the presence of many different alleles of the causal gene (multiple colours) makes naive GWA scans (such as in **a**) ineffective. **c**, GWA scans miss rarely occurring alleles that have a large effect (the top locus) but locate common alleles with smaller effects (the centre locus). In all three cases (**a**–**c**), the causal loci can be identified in experimental crosses (shaded areas) by using linkage mapping instead of GWA mapping, but alleles identified in this way might be rare in the population and therefore not informative for most individuals and not representative of the population.

Future directions

In the past two years, numerous GWA studies of humans have been published[29]. Despite the success of these studies, the identified alleles typically provide an explanation for only a small proportion of the variation in the population[30]. Similar studies in plants, which can be expected to be published in the next two years, will tell the genetics community whether alleles with large effects are more common in plants than was found in human GWA studies.

A serious deficiency in the understanding of plant evolution stems from the inability to compare the genomes of close relatives. The immense value of such studies has been aptly demonstrated in yeasts, *Caenorhabditis*, *Drosophila* and primates. Comparisons of recently diverged genomes allow not only the mapping of conserved genomic elements but also the detection of lineage-specific selection or the identification of recently introgressed segments from relatives. The last is of particular interest in plants, for which interspecific hybrids can often give rise to viable offspring. The closest relative of *A. thaliana* for which a genome sequence is available is papaya, which diverged from *A. thaliana* 70 million years (Myr) ago[31]. This lack of appropriate genome sequence information is ironic, because for years there has been more information about within-species polymorphism in plants than in other non-human organisms[14,15,32]. The plant genetics community is therefore excited that the sequencing of the genomes of *Arabidopsis lyrata* and *Capsella rubella* — close relatives of *A. thaliana* that diverged only about 5 Myr ago and 10 Myr ago, respectively — is almost complete. Similarly, sequencing the genomes of wild progenitors of rice holds the promise of providing important information about the domestication of rice.

Despite the excitement about studying variation between species and between genera, the most revealing insights are likely to come from much more detailed studies of within-species sequence diversity. Until recently, such analyses have focused on relatively minor differences between individuals, such as SNPs, and small insertions and deletions. It has become increasingly clear, however, that individuals, whether maize, *A. thaliana* or humans, can differ by the presence of hundreds of genes and by large genomic rearrangements[1,5,33–35].

One such in-depth study of within-species diversity is the recently launched '1000 Genomes' project (http://1000genomes.org), which seeks to generate a deep catalogue of genetic variation by sequencing the genomes of at least 1,000 individuals from around the world, using next-generation sequencing technologies. We are advocates of a complementary project for *A. thaliana*: 1001 Genomes (http://1001genomes.org). *A. thaliana* is almost uniquely suited to such an effort, given the advantages it offers for analysing population genetics and studying genotype–environment interactions. First, *A. thaliana* genomes are less than one-twentieth the size of human genomes and are much less repetitive in structure, making them less expensive to sequence and much easier to reassemble than human genomes. Second, the genomes of most *A. thaliana* accessions (or strains) are naturally inbred, whereas humans (and most model animals) are heterozygous. Therefore, when using a whole-genome shotgun strategy to sequence *A. thaliana* genomes, the problems of determining haplotype phase (that is, the arrangement of alleles on homologous chromosomes) are avoided, greatly simplifying downstream analyses. Last, for practical and ethical reasons, it will not be possible to use the individuals of the 1000 Genomes project directly in GWA studies, with the exception of measuring phenotypes for cell lines derived from samples from the 1,000 individuals. Instead, the information will need to be used as a resource to impute genotypes in other samples. By contrast, an endless supply of seeds can be produced for any naturally inbred *A. thaliana* accession that is chosen for genome sequencing. Therefore, any number of plants with an identical genotype can be grown for each accession and then phenotyped in as many environments as desired. So the sequence information that is collected can be used directly in GWA studies at many levels, including at the biochemical, metabolic, physiological, morphological and whole-plant-fitness levels.

With the combination of GWA studies and forward-genetic approaches, it will finally become possible to bridge the genotype–phenotype divide, at least in *A. thaliana*. And, as similar projects are set up to study other species, it will become possible to answer general questions about the molecular genetic basis of evolutionary change. ■

1. Nilsson-Ehle, H. Kreuzungsuntersuchungen an Hafer und Weizen. 1. *Lunds Universitets Årsskrift* **5,** 1–122 (1909).
2. East, E. M. A Mendelian interpretation of variation that is apparently continuous. *Am. Nat.* **44,** 65–82 (1910).
3. Sax, K. The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* **8,** 552–560 (1923).
4. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386,** 485–488 (1997).
5. Frary, A. *et al. fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289,** 85–88 (2000).
6. Fridman, E., Pleban, T. & Zamir, D. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl Acad. Sci. USA* **97,** 4718–4723 (2000).
This study pinpointed a QTL to a small intragenic region by using recombination mapping.
7. Yano, M. *et al. Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering-time gene *CONSTANS*. *Plant Cell* **12,** 2473–2483 (2000).
8. Bentsink, L., Jowett, J., Hanhart, C. J. & Koornneef, M. Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **103,** 17042–17047 (2006).

9. Mouchel, C. F., Briggs, G. C. & Hardtke, C. S. Natural genetic variation in *Arabidopsis* identifies *BREVIS RADIX*, a novel regulator of cell proliferation and elongation in the root. *Genes Dev.* **18,** 700–714 (2004).
10. Macquet, A. *et al.* A naturally occurring mutation in an *Arabidopsis* accession affects a β-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage. *Plant Cell* **19,** 3990–4006 (2007).
11. Baxter, I. *et al.* Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (*MOT1*). *PLoS Genet.* **4,** e1000004 (2008).
12. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
13. Frazer, K. A. *et al.* A second-generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
14. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3,** e196 (2005).
15. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317,** 338–342 (2007).
This study provided the first genomic view of the pattern of polymorphism in a plant.
16. Kim, S. *et al*. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.* **39,** 1151–1155 (2007).
17. Shindo C. *et al.* Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of *Arabidopsis*. *Plant Physiol.* **138,** 1163–1173 (2005).
18. Thornsberry, J. M. *et al. Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genet.* **28,** 286–289 (2001).
19. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* **44,** 1054–1064 (2005).
20. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38,** 303–308 (2006).
This study identified an effective way to account for population structure in association mapping, an innovative approach that has had a large impact on the field.
21. Aranzana, M. J. *et al*. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* **1,** e60 (2005).
22. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3,** e4 (2007).
23. Harjes, C. E. *et al.* Natural genetic variation in *lycopene epsilon cyclase* tapped for maize biofortification. *Science* **319,** 330–333 (2008).
24. Balasubramanian, S. *et al.* The *PHYTOCHROME C* photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. *Nature Genet.* **38,** 711–715 (2006).
25. Ewens, W. & Spielman, R. The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.* **57,** 455–464 (1995).
26. Yu, J., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178,** 539–551 (2008).
This report introduced an elegant experimental combination that simultaneously exploits the advantages of mapping in experimental crosses and the increased resolution of association mapping.
27. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38,** 879–887 (2006).
28. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36,** 1133–1137 (2004).
29. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
30. Iles, M. M. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.* **4,** e33 (2008).
31. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452,** 991–996 (2008).
32. Wright, S. I. The effects of artificial selection on the maize genome. *Science* **308,** 1310–1314 (2005).
33. Buckler, E. S., Gaut, B. S. & McMullen, M. D. Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* **9,** 172–176 (2006).
34. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
35. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).

Acknowledgements Studies of natural genetic variation are supported by the National Science Foundation and the National Institutes of Health (M.N.), and by the German Research Foundation, the German Federal Ministry of Education and Research, the European Union's Sixth Framework Programme, the Human Frontier Science Program and the Max Planck Society (D.W.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (magnus@usc.edu; weigel@weigelworld.org).

NATURE|Vol 456|11 December 2008|doi:10.1038/nature07630

Applying mouse complex-trait resources to behavioural genetics

Jonathan Flint[1] & Richard Mott[1]

Studies of the genetic basis of behaviour in mice are at a turning point. Soon, new resources will enable the behavioural function of all genes to be tested and the networks of genes, messenger RNAs and proteins involved in a particular behaviour to be identified. Using these resources, scientists will be able to analyse mouse behaviour at an unprecedented level of detail. Interpreting the new data, however, will require a shift in focus from gene-based approaches to network-based approaches.

In the 1990s, technological developments in molecular biology presented mouse behavioural geneticists with two new genetic strategies. The first strategy was to use targeted mutations to identify genes involved in abnormalities in learning and memory[1,2]. More specifically, certain genes were knocked out by the introduction of mutations by homologous recombination in mouse embryonic stem cells, and then lines of mice were bred from these cells and studied in behavioural assays. This approach is powerful but technically complex and expensive, so its use was limited to investigating the function of cloned genes that had already been implicated in behavioural processes. The second strategy, the mapping of natural genetic variation that affects mouse behaviour, held the promise of establishing the chromosomal location of genes involved in almost any measurable behaviour[3]. Thus, a combination of these two strategies emerged as an approach for identifying the genes involved in mouse behaviour. The idea was that, first, genetic association mapping (the latter strategy) would be carried out in crosses between inbred mouse strains to determine the approximate location of genes that affect a particular behaviour. Then, after further genetic and molecular analysis had identified a small number of candidate genes, each gene in the region would be knocked out. Behavioural testing of gene-knockout mice would be used to determine the behavioural function of the deleted gene. And a variant of complementation testing (in which the mutant mouse is crossed to inbred strains carrying allelic variants at the locus)[4] could be used to confirm the gene's role at the locus and the behaviour it influenced[5]. Combining association mapping and gene-knockout technologies therefore seemed to be the way forward for understanding the biology of behaviour in mice.

Progress soon stalled, however — on both sides of the strategy. It became evident that the effect of a targeted mutation on behaviour was not straightforward. Phenotypes varied according to the genetic background of the inbred strain on which the knockout was placed[6], and mutations affecting behaviour also affected other phenotypes[7,8]. Nevertheless, targeted mutations became one of the standard tools for behavioural analysis. Association mapping met a worse fate. When attempts were made to find candidate genes at a region identified by mapping, that single locus frequently fractionated into numerous smaller loci[9]. In addition, in some cases, evidence emerged that the phenotypic consequences of one locus depended on whether a particular allele was present at another locus (a phenomenon known as epistasis)[10]. Consequently, the identification of genes associated with particular behavioural traits has been slow[5].

In this Progress article, we describe the resources and methods that are being used for studying the genetic basis of behaviour on a genome-wide scale. We then discuss the challenges that confront mouse behavioural genetics and explain how new resources (a knockout mouse for every gene, and mice designed for high-resolution mapping), together with new methods (such as network analysis of transcripts), are being mobilized to tackle these issues. We argue that using these new resources for the analysis of complex traits necessitates a shift in focus away from individual genes and towards networks of genetic and molecular interactions.

Methods and resources available now

Three additional strategies are available to find the genes involved in mouse behaviour, complementing the now classic methods of mapping and gene-knockout analyses: these are mutagenesis, *in silico* mapping and transcriptome analysis. We outline these approaches and then consider why they have not yet had a major impact on gene discovery.

First, on the assumption that the effect of mutations on behaviour can be large, a mutagenic compound, *N*-ethyl-*N*-nitrosourea (ENU), can be used to induce mutations in mice, a strategy that has successfully been used to analyse *Drosophila melanogaster* behaviour[11]. It was hoped that it would prove easy to identify mutations by screening small pedigrees for abnormal behaviour. These mutations would then be cloned, thereby revealing the identity of the affected genes[12]. Mutagenesis has worked for identifying genes that affect circadian rhythm[13], but it has not been as successful for other behaviours, mainly because the changes in a mutant animal are too subtle to be easily detected in the small pedigrees that are routinely screened.

Second, a computational method can be used to predict the genes associated with particular mouse behaviours, on the basis of DNA sequence variation between inbred laboratory strains and the known behavioural phenotypes of these mice[14]. Most inbred mouse strains originated from a limited population of mice in just a few laboratories, and alignment of the genomic sequences shows that the genomes of these mice are mosaics, with most segments derived from only a few sources[15]. This mosaicism can be exploited to carry out the equivalent of a genome-wide association study using published genotypes and phenotypes of inbred strains (thus the method can be carried out without genotyping and is known as *in silico* mapping). This method has proved effective for localizing large genetic effects but not, so far, for identifying the small effects that characterize behavioural genes. This is because genetic relatedness between the genomes of inbred mouse strains makes the method difficult to apply[16].

Third, combining genetic data with transcript (or gene-expression) data has been proposed as a way to find genes in cases in which behavioural

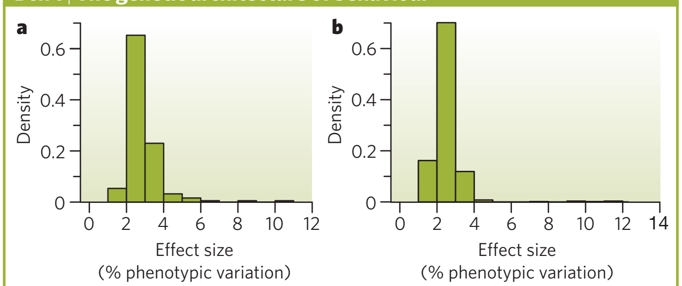
[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.

variation results from transcript abundance[17]. Identifying loci that cause both transcript variation and phenotypic variation can help to find the genes responsible, a strategy that has paid off in studies of several complex diseases[18]. Rather than seeking to identify the sequence variants that underlie phenotypic variation, correlations between variations in transcript abundance can be used to identify gene-expression networks that, when perturbed by the variation at susceptibility loci, contribute to phenotypic variation[19]. However, so far, the approach has had limited application[20].

Why have these approaches not been productive? The single biggest factor preventing the identification of genes that affect behaviour is probably that, as mentioned earlier, each sequence polymorphism contributes only a small amount to behavioural variation, even when a polymorphism results in the inactivation of a gene. And, in this respect, behaviour as a phenotype is no different from other phenotypes (Box 1). The extent to which a particular locus contributes to behaviour is typically reported as effect size — that is, the percentage of phenotypic variation that is attributable to allelic variation at a locus. The effect sizes of naturally occurring variants depend on the frequency with which these alleles occur in the population and therefore on the population in which they are mapped. These effect sizes are usually small. For example, in a cross between two inbred mouse strains, a detectable locus contributes to about 5% of the variation in a behavioural phenotype[21]. By contrast, in a complex quasi-outbred stock known as the heterogeneous stock (HS, which was derived from eight inbred mouse strains that were crossbred for about 60 generations), a locus contributes to about 3% of the behavioural variation[22]. And, in a fully outbred population, the effect size of a particular locus might account for much less than 1% of the phenotypic variation (the contribution attributable to each locus will be less because the mice have more genetic variability than the quasi-outbred stock).

It is less widely recognized that knocking out a gene also has a relatively small impact on behaviour. As far as we know, no one has measured the effect on behaviour of a gene knockout segregating in a cross, but this effect can be estimated from reported means and standard deviations of behavioural parameters for gene-knockout mice and wild-type mice. For example, using data published on the phenotypic effect of knocking out *Crhr2* (which encodes corticotropin-releasing hormone receptor 2)[23], the knocked-out allele accounts for about 10% of the total phenotypic variation. This is only twice the effect size that would be expected for a naturally occurring sequence variant and almost certainly explains why it has been difficult to identify genes that affect behaviour by using the strategy of ENU-induced mutagenesis[24].

Box 1 | The genetic architecture of behaviour



It is now known that, from a genetic perspective, there is nothing special about behaviour, at least in mice. Quantitative trait loci (QTLs) have been mapped in heterogeneous stock (HS) mice, the only resource for which high-resolution association mapping has been carried out for multiple phenotypes[22]. For 97 phenotypes, 843 QTLs were mapped. The effect sizes are shown for 187 behavioural QTLs (see figure, panel **a**) and 656 non-behavioural QTLs (panel **b**), with effect size plotted against density (the proportion of QTLs displaying that effect size). These two groups of QTLs are almost identical in terms of the distribution of their effect sizes, contradicting the view that the genetic analysis of behaviour would be much easier if phenotypic assays for behaviour could be improved or if the phenotypes under study were more genetically tractable (for example, if components of complex behavioural phenotypes, sometimes called endophenotypes, were studied).

Table 1 | Phenotypes of gene-knockout mice from Lexicon and Deltagen

Phenotype studied	Number of strains with an abnormal phenotype	Percentage of strains with an abnormal phenotype	Total number of strains tested
Open-field activity	39	19	206
Albumin	18	7	254
Alkaline phosphatase	18	7	254
Blood urea nitrogen	23	9	254
Calcium	16	6	254
Cholesterol	26	10	254
Glucose	10	4	254
Haemoglobin	11	5	243
Leukocytes	15	6	243
Phosphorus	18	7	254
Platelets	14	6	243
Triacylglycerols	17	7	254

A survey of about 250 targeted knockouts generated by two private companies (Lexicon Pharmaceuticals and Deltagen) is shown. Data were obtained from the Mouse Genome Informatics database (http://www.informatics.jax.org/external/ko). The phenotype of each gene-knockout mouse was compared with that of wild-type mice, and the number, and the percentage of the total number, of gene-knockout strains that showed a significant difference ($P<0.05$ in a t-test) from wild-type mice is indicated. Complete phenotypic information was not available for all mice, so the total number of mutant mice tested varies between phenotypes.

New resources

In the next few years, several new resources are going to be placed into the hands of mouse geneticists, helping to surmount these challenges. Surprisingly, two of these are a reprise of the classic duo: mutants (a complete set of gene knockouts) and mapping (a large set of inbred strains known as the Collaborative Cross).

Knockouts will be made available through the International Mouse Knockout Consortium, which was launched in 2007. As part of this consortium, researchers involved in four programmes are collaborating to mutate all protein-coding genes in mice, using a combination of gene trapping (a way of introducing insertional mutations across the mouse genome) and gene targeting in mouse embryonic stem cells[25]. The plan is to generate knockouts of all protein-coding genes by 2010. In addition, transposon-based resources are on the horizon[26]: mutants will be generated by inserting a known sequence so that, unlike for ENU-induced mutants, the mutated genes can be identified without positional cloning.

The availability of a gene-knockout mouse corresponding to every protein-coding gene in the genome opens up the possibility of measuring the behavioural phenotype of all of these mouse strains. A glimpse of what this might uncover comes from phenotypic data obtained from about 250 gene-knockout strains that were acquired by the National Institutes of Health from two companies (Lexicon Pharmaceuticals and Deltagen) and now form part of several public repositories; the phenotypic data are available from the Mouse Genome Informatics database (http://www.informatics.jax.org/external/ko). Although these strains are not an unbiased selection of mutants, data from these mice are valuable because this is the only instance in which a large number of gene-knockout mice have been subjected to a broad range of measurements. Table 1 shows the percentage of mice with fear-related behaviour (activity in an open-field arena) and the percentage of mice with biochemical and haematological abnormalities. Almost 20% of the mouse strains showed a significant change in behavioural phenotype compared with littermate controls.

On the one hand, finding a behavioural phenotype in such a large proportion of the gene-knockout mice suggests that phenotyping all available mutants would be productive, even if the effects of mutations on open-field activity phenotype arise from outside the central nervous system and therefore indicate little about the biology of fear-related behaviour. Indeed, this finding suggests that a systematic gene-knockout programme could supersede association mapping strategies as a way of finding the genes that are involved in behaviour. One argument against a systematic gene-knockout programme, however, is the cost: about 20,000 mouse lines would need to be generated, and then multiple animals from each line would need to be tested for each behaviour of interest. Even

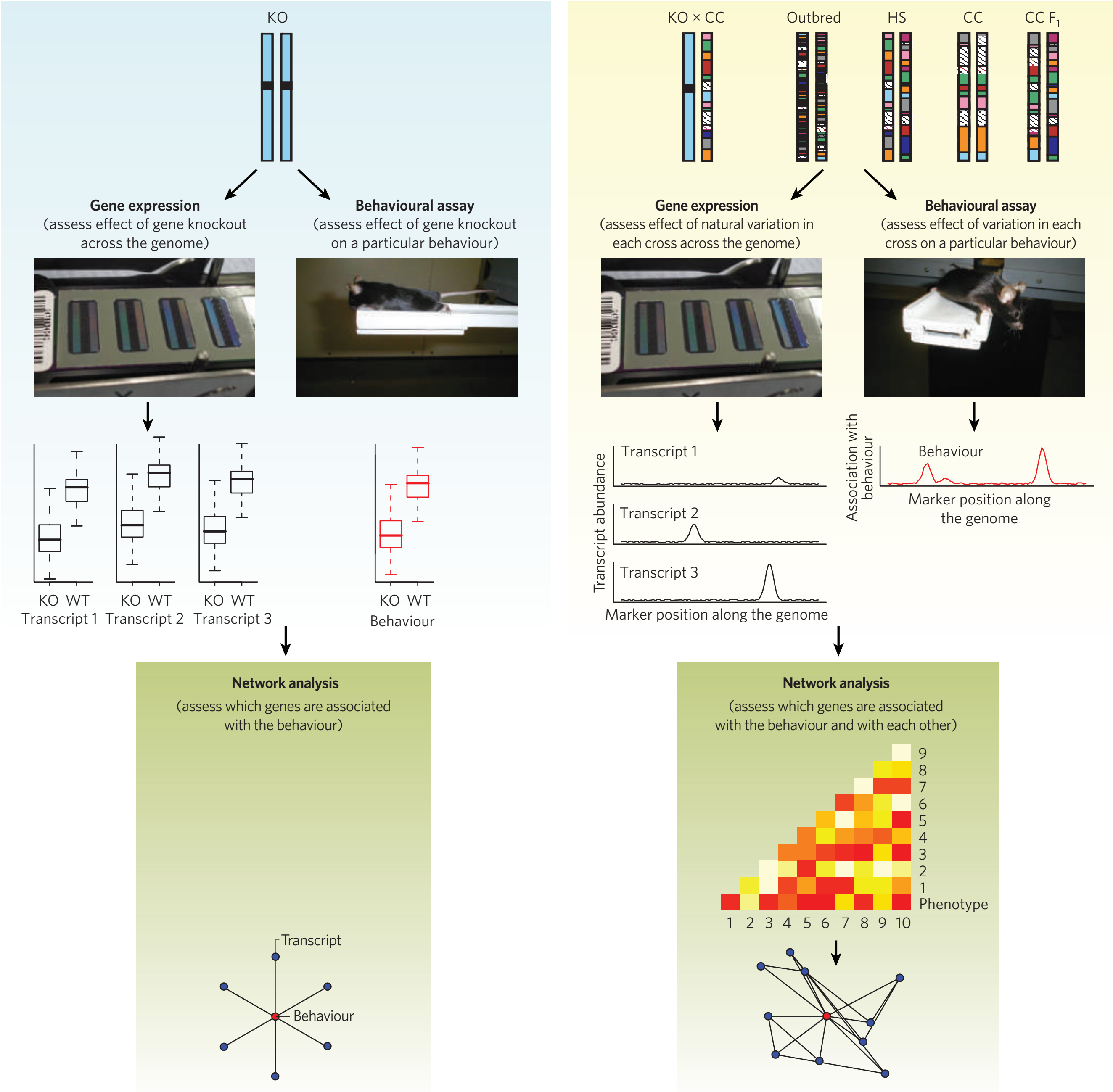


Figure 1 | Genetic dissection of behaviour. Strategies based on gene knockouts (KO; left panel) and population-based genetics (right panel) are shown. The various mouse stocks available as resources for population-based studies include outbred mice, Collaborative Cross (CC) mice and heterogeneous stock (HS) mice. Various types of experiment are common to both approaches; the two experiments depicted are behavioural assays and gene-expression assays. Each pairing of experiment and resource needs to be analysed in a specific way, and the data can then be integrated in a network analysis. A typical genome of each resource is depicted (top) by a pair of chromosomes, coloured to indicate the degree of mosaicism in terms of standard laboratory strains. A KO mouse genome (left) contains a single genetic alteration (black), usually on a fixed wild-type (WT) genetic background (blue). Crossing a non-recessive KO mouse with a CC strain (KO × CC) produces mice suited for mapping genetic modifiers of the knocked-out gene. In terms of gene expression (black plots), the relative abundance of many thousands of transcripts is measured in a tissue relevant to behaviour (such as the hippocampus). Using the KO mouse strategy, transcripts present at significantly different levels in KO mice and WT mice are implicated (data for three transcripts are shown). Using the population-based strategy, gene expression is assayed across the genome for a group of mice from each resource (representing the population), and the regions where differences in DNA sequence correlate with differences in gene expression (known as expression quantitative trait loci, eQTLs) are mapped (one is shown per plot). In terms of behaviour (red plots), using the KO strategy, a gene is implicated if there are significant differences in the behavioural parameter between KO mice and WT mice. Using the population-based strategy, the population-wide variation in behaviour between mice is studied and used to map the genetic loci where sequence differences and behavioural differences are correlated (red plot shows two QTLs). Both types of strategy produce data that can be viewed as genetic network analyses. For population-based studies, the abundances of certain transcripts are first correlated (depicted as the plot of coloured squares, with the strength of the correlation reflected by the colour intensity from yellow to red and with correlations for ten transcripts shown). Then, the stronger correlations can be used to define a gene network, into which data about the behavioural trait can be embedded to identify genes whose expression correlates with the phenotype (lines are drawn between correlated transcripts (blue), and transcripts whose abundance correlates with behavioural variation are linked to the behaviour node (red)). In KO studies, the network is simpler, comprising the phenotypes and the genes with significant changes in expression. In either case, the network can be annotated further: for example, by labelling transcript nodes with eQTLs that co-localize with the behavioural QTLs.

maintaining such a large population of mice would be challenging, but it is likely to be feasible if the economies of scale are sufficient.

On the other hand, implicating 20% of the genes in the genome in fear-related behaviour might not be useful. Even if only 5% of these genes are shown to affect the central nervous system, it is unclear how enlightening it will be to have identified these genes. As an example, about a decade ago, Joshua Sanes and Jeff Lichtman observed that about 100 genes had been implicated in long-term potentiation[27], a cellular phenomenon that is thought to underlie memory processes in the hippocampus. Yet despite the knowledge that these 100 genes are involved, an accepted molecular explanation of long-term potentiation has not emerged. It has become clear that explaining the way in which many, apparently unrelated, molecules are involved in a complex phenomenon is difficult when following the causal model that researchers adopted from biochemical genetics: that is, that genes encode molecules that interact in linear pathways. A systematic gene-knockout project for investigating the effects of genes on behaviour provides information about one gene at a time. This approach is likely to be inefficient at working out how behaviour arises from an individual's genetic make-up, partly because it needs to proceed gene by gene and partly because interactions between genes might be an important contributor.

For this reason, it will be valuable to have a systems-biology resource to correlate behavioural variation with variation in sequence, together with variation in transcript, protein and metabolite abundance (Fig. 1). Genome-wide studies of interactions between genes[28] and between proteins[29] in single-celled organisms (yeast) have already proved that it is possible to identify structure within complex biological networks. Similar approaches are possible in mice[19,30]. For example, analysis of the proteins that interact with just one synaptic component (the NMDA receptor) in the mouse brain has shown the potential of this approach for identifying key molecules in the biology of learning and memory[31].

Data on gene, transcript and protein interactions might one day become available by assaying knockout mice corresponding to each protein-coding gene, but it will be more efficient to obtain this information by studying a large set of recombinant inbred strains. One such group of strains is being developed: known as the Collaborative Cross[32], it will consist of 1,000 mouse strains descended from 8 genetically diverse founder strains, and the new strains are due to be fully inbred by 2012. The clonal nature of the Collaborative Cross strains means that multiple assays will be able to be carried out on each genome; for example, it will be possible to determine transcript abundance in every tissue at every stage of development, as well as to study the effects of multiple environmental interventions. In addition, the first-generation progeny of Collaborative Cross strains will form a set of mouse strains with hundreds of thousands of reproducible heterozygous genotypes.

There is one drawback to the Collaborative Cross, however: it will not deliver gene-level resolution[33]. Association mapping data could be obtained at a very high level of resolution by using wild mice[34], but because of the genetic structure of this outbred population, similar to an outbred human population, tens of thousands of individuals would need to be analysed in order to uncover how the genes were influencing behaviour[35].

One untapped resource that might help to solve the problem of resolution is commercially available outbred populations. Some of these populations have been maintained in very large colonies for many generations and have a density of accumulated recombination similar to that in human populations. These populations are not as genetically diverse as completely wild populations, and a few hundred animals can be used for association mapping at single-gene resolution[5]. Furthermore, the Collaborative Cross itself could be used to generate an outbred cross, delivering a higher-resolution mapping resource to complement the large set of reproducible genotypes.

Perspective

Finding the genes that are involved in behaviour remains a challenging endeavour, although it is about to become easier. But even complete lists of genes that affect behaviour, and knowledge of the exact molecular changes that give rise to behaviour, will not lead to an understanding of how behaviour emerges from the biological 'substrate'. For this, researchers will need to investigate behaviour at multiple levels (at the level of the gene, the transcript and the protein), which will be made possible by the resources discussed here. These complex-trait resources will provide an unparalleled opportunity to pull apart the biology of behaviour. ■

1. Silva, A. J., Paylor, R., Wehner, J. M. & Tonegawa, S. Impaired spatial learning in α-calcium-calmodulin kinase II mutant mice. *Science* **257,** 206–211 (1992).
2. Grant, S. G. N. *et al.* Impaired long-term potentiation, spatial learning and hippocampal development in *Fyn* mutant mice. Science **258,** 1903–1910 (1992).
3. Flint, J. *et al.* A simple genetic basis for a complex psychological trait in laboratory mice. *Science* **269,** 1432–1435 (1995).
4. Long, A. D., Mullaney, S. L., Mackay, T. F. C. & Langley, C. H. Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*. *Genetics* **144,** 1497–1510 (1996).
5. Yalcin, B. *et al.* Genetic dissection of a behavioral quantitative trait locus shows that *Rgs2* modulates anxiety in mice. *Nature Genet.* **36,** 1197–1202 (2004).
This was the first report of using quantitative complementation to identify the effect of a quantitative trait locus gene on behaviour.
6. Silva, A. J. *et al.* Mutant mice and neuroscience: recommendations concerning genetic background. *Neuron* **19,** 755–759 (1997).
7. Amieux, P. S. *et al.* Increased basal cAMP-dependent protein kinase activity inhibits the formation of mesoderm-derived structures in the developing mouse embryo. *J. Biol. Chem.* **277,** 27294–27304 (2002).
8. Huang, Y. Y. *et al.* A genetic test of the effects of mutations in PKA on mossy fiber LTP and its relation to spatial and contextual learning. *Cell* **83,** 1211–1222 (1995).
9. Talbot, C. J. *et al.* High-resolution mapping of quantitative trait loci in outbred mice. *Nature Genet.* **21,** 305–308 (1999).
10. Shimomura, K. *et al.* Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.* **11,** 959–980 (2001).
11. Tully, T. *et al.* A return to genetic dissection of memory in *Drosophila*. *Cold Spring Harb. Symp. Quant. Biol.* **61,** 207–218 (1996).
12. Nadeau, J. H. & Frankel, W. N. The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. *Nature Genet.* **25,** 381–384 (2000).
13. Vitaterna, M. H. *et al.* Mutagenesis and mapping of a mouse gene, *Clock*, essential for circadian behaviour. *Science* **264,** 719–725 (1994).
14. Grupe, A. *et al.* *In silico* mapping of complex disease-related traits in mice. *Science* **292,** 1915–1918 (2001).
15. Wade, C. M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420,** 574–578 (2002).
16. Payseur, B. A. & Place, M. Prospects for association mapping in classical inbred mouse strains. *Genetics* **175,** 1999–2008 (2007).
17. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
This paper introduced a method to investigate causal relationships between sequence variants, gene expression and phenotypes.
18. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genet.* **37,** 1224–1233 (2005).
19. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
20. Hovatta, I. *et al.* Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature* **438,** 662–666 (2005).
21. Flint, J., Valdar, W., Shifman, S. & Mott, R. Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Rev. Genet.* **6,** 271–286 (2005).
22. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38,** 879–887 (2006).
This study shows that high-resolution genome-wide association mapping of behaviour, and other traits, is possible in outbred mice.
23. Kishimoto, T. *et al.* Deletion of *Crhr2* reveals an anxiolytic role for corticotropin-releasing hormone receptor-2. *Nature Genet.* **24,** 415–419 (2000).
24. Keays, D. A. *et al.* Mutations in α-tubulin cause abnormal neuronal migration in mice and lissencephaly in humans. *Cell* **128,** 45–57 (2007).
25. Collins, F. S., Rossant, J. & Wurst, W. A mouse for all reasons. *Cell* **128,** 9–13 (2007).
26. Ding, S. *et al.* Efficient transposition of the *piggyBac* (*PB*) transposon in mammalian cells and mice. *Cell* **122,** 473–483 (2005).
27. Sanes, J. R. & Lichtman, J. W. Can molecules explain long-term potentiation? *Nature Neurosci.* **2,** 597–604 (1999).
28. Tong, A. H. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813 (2004).
29. Wuchty, S., Oltvai, Z. N. & Barabasi, A. L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genet.* **35,** 176–179 (2003).
30. Keller, M. P. *et al.* A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* **18,** 706–716 (2008).
31. Choudhary, J. & Grant, S. G. Proteomics in postgenomic neuroscience: the end of the beginning. *Nature Neurosci.* **7,** 440–445 (2004).
32. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36,** 1133–1137 (2004).
33. Valdar, W., Flint, J. & Mott, R. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* **172,** 1783–1797 (2006).
34. Laurie, C. C. *et al.* Linkage disequilibrium in wild mice. *PLoS Genet.* **3,** e144 (2007).
35. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638–645 (2008).

Acknowledgements We thank the Wellcome Trust for support and are grateful to S. McCormick for comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to J.F. (jf@well.ox.ac.uk).

NATURE|Vol 456|11 December 2008|doi:10.1038/nature07631

Progress and challenges in genome-wide association studies in humans

Peter Donnelly

After more than a decade of hope and hype, researchers are finally making inroads into understanding the genetic basis of many common human diseases. The use of genome-wide association studies has broken the logjam, enabling genetic variants at specific loci to be associated with particular diseases. Genetic association data are now providing new routes to understanding the aetiology of disease, as well as new footholds on the long and difficult path to better treatment and disease prevention.

Genome-wide association (GWA) studies look for associations between DNA sequence variants and phenotypes of interest. They do so by studying individuals with different phenotypes (for example, affected or not affected by a common complex disease) and determining their genotype at the positions of hundreds of thousands of single nucleotide polymorphisms (SNPs). SNPs for which one variant is statistically more common in individuals belonging to one phenotypic group are then reported as being associated with the phenotype. A subsequent experiment would seek to confirm the association by examining whether the original results are replicated in a new sample of individuals.

More than 300 replicated associations have now been reported for more than 70 common diseases, conditions and biological measurements as a result of GWA studies (see the National Human Genome Research Institute catalogue of published GWA studies, http://www.genome.gov/gwastudies). As an example, there are now more than 30 loci known to be associated with the autoimmune disease Crohn's disease[1], almost 20 loci associated with type 2 diabetes[2] and more than 40 loci associated with height[3,4]. Just two years ago, the corresponding numbers were two, three and zero, respectively.

In a GWA study, the ability to detect an association at a particular SNP decreases with the frequency of the less common variant at that SNP. Thus, studies have so far been better powered to detect associations with more common variants, and this is reflected in their findings, with few of the reported associations involving rare variants. Another common feature of variants found in GWA studies is that their effect on disease risk is small: typically, each additional copy of the variant increases disease risk by 10–30%.

With a few exceptions, details of the functional mechanisms responsible for the association signals are not yet known. Nonetheless, there have been some exciting insights into disease processes, as well as tantalizing glimpses of unsuspected overlaps between disease loci[5,6] (Box 1). Strikingly, most of the loci identified so far did not feature on lists of candidates for genes involved in the disease in question.

Yet the findings so far are just the tip of the iceberg. It is clear that more extensive analyses, follow-up studies and meta-analyses of published studies will uncover more associated variants. In addition, many new GWA studies are planned or are underway. For example, the Wellcome Trust (a major charity that funds medical research in the United Kingdom) has committed more than £30 million ($45 million) to studies of DNA samples from 120,000 individuals across 27 phenotypes and has just issued a call for further grant applications in this area. Given the level of investment in these studies, and the extent of their success so far, it is important to examine the challenges that lie ahead for researchers in better understanding and using the findings from GWA studies.

Challenges

One challenge for the next wave of GWA studies arises because companies that provide high-throughput genotyping on a contract basis have brought GWA studies within the reach of clinical research groups. To begin a GWA study, all that is needed is DNA from a large sample of patients (appropriate control genotypes might already be publicly available) and a successful grant application. There is no need to have a genotyping kit or expertise in genomics or data analysis. But there is a crucial caveat: good quality control is essential, because there are myriad artefacts that might easily be mistaken for real genetic effects. The success of GWA studies in providing results that are reliable and able to be replicated owes a great deal to the careful attention paid to quality control in the early studies, and it is to be hoped that this tradition is not lost as the approach is applied much more widely.

In another sense though, researchers had it too easy in the early GWA studies. A typical outcome of these studies was one or several 'hits' with very strong statistical support. The next natural step was to assess whether these associations held for other samples, and these replication studies were also often successful. It is now clear, however, that 'true' associations also lurk much lower down a ranked list of hits. For some diseases, researchers have now worked down the list as far as some of these lower-ranked associations, often after carrying out meta-analyses and extensive follow-up studies. But now might be the time for a change of approach. Instead of initially following up only the top few hits, it seems more efficient to take advantage of economies of scale in genotyping and undertake a single experiment to follow up many of the top hits (for example, 1,536, a natural quantum for one common platform, or even 5,000–10,000) in large replication samples, as was recently reported for a GWA study of coeliac disease[7]. The size of the initial study is still important though, because true positive hits will tend to be higher up the list of putative associations in a large initial study and therefore are more likely to be followed up in replication studies. Furthermore, large initial studies, in contrast to some multiphase study designs, also provide a powerful resource for meta-analyses.

The associations uncovered by GWA studies (and confirmed by appropriate replication studies) are between sets of SNPs in small genomic regions (typically 50–250 kilobases, but in some cases more than 1 megabase) and phenotypic variation. So another point to consider is that the disease-associated SNPs that have been uncovered so far are unlikely to be the functional variants. Instead, they act as proxies, or sentinels, for the true disease variant(s). Because of their shared evolutionary history, alleles at nearby SNPs on the same human chromosome are often correlated: chromosomes that carry an 'A' allele at one SNP might often carry a 'C' allele at a nearby SNP, and so on. This phenomenon, known as linkage

disequilibrium, means that it is not essential to type the actual causal variant in an association study. Genotypes at the SNPs assayed in the study will often be correlated with those at the causal variant(s), and, provided that the study is large enough, there will still be an association signal at these correlated SNPs (or proxies). Therefore, one of the major challenges now is to 'drill down' into the associated regions to define the causal variants and to uncover how they contribute to disease.

One natural follow-on strategy, fine mapping, involves genotyping at the positions of many more SNPs in the associated region than in the original GWA study, and in large numbers of individuals with and without the disease of interest (cases and controls). If a causal SNP is genotyped in the fine mapping, its association signal should be stronger than that of many of the other SNPs that are genotyped. Ideally, fine mapping would involve genotyping as many of the correlated SNPs as possible in the associated region. Therefore, before fine mapping is carried out, it might be profitable to resequence the associated region in a set of cases and controls, in order to identify much of the SNP variation that is present, although the rapidly increasing catalogue of known variation (being put together by the 1000 Genomes Project, http://www.1000genomes.org) is reducing the need for targeted resequencing, at least in controls. Few such fine-mapping experiments of associated regions have been published, and there are important open questions about the best design for such studies. No genotyping experiment will be able to distinguish between a causal SNP and other SNPs that are perfectly correlated with this SNP. Therefore, even when it is successful, fine mapping is unlikely to point to just one potentially causal SNP and, instead, will typically narrow researchers' attention to a set of such SNPs, which would then need to be studied in functional assays.

This stage will be difficult. The familiar paradigm from the many Mendelian human diseases whose genetics is now understood is that the functional mutations act through truncating or altering gene products. This is also typical of the relatively few known examples of rare mutations with large or intermediate effects on susceptibility to common human diseases. By contrast, there is growing evidence that a sizeable proportion, perhaps most, of the functional variants that underlie associations in GWA studies exert their effects through gene regulation rather than changing gene products. For example, several associated regions found in GWA studies do not contain any known genes. In terms of appropriate functional studies for assessing GWA findings, the extent to which familiar paradigms developed in model systems for Mendelian conditions, such as gene knockouts, will be useful for understanding the more subtle landscape of regulatory changes emerging from GWA studies is unclear at present. This adds to the pressing demand for progress towards a deeper understanding of the vast regulatory regions within genomes, as well as for tools in model systems that will aid in this effort.

An alternative way to follow up GWA findings is not to search for the variant responsible for the association signal but, instead, to borrow a strategy from Mendelian genetics and screen the genes in the associated regions for rare mutations. If mutations in such a gene occur statistically more frequently in the cases than in the controls, this could implicate the gene in the disease process, regardless of whether it is responsible for the original association signal. Like fine mapping, there are open questions about the design and efficacy of this approach.

Predicting risk

In addition to increasing the understanding of disease pathophysiology and aetiology, findings from GWA studies might also allow predictions of an individual's risk of developing a disease to be refined. However, although it seems inevitable that genetic information will be used routinely as part of many aspects of medical care in the medium term, there is still much that is unknown about how to apply this information.

One issue is that the risks associated with the presence of a causal variant in a known region of association are unclear, and are likely to be underestimated at present, because most causal variants have not been identified. At present, risk is assessed by assaying an individual's genotype at the position of the proxy SNP uncovered by GWA analysis. Because this is a noisy measure of the genotype at the position of the causal SNP, an individual's risk of disease is typically underestimated. For the same reason, the statistical power to determine the correct disease model, for example whether the effect is recessive or dominant, is low unless the true variants, or variants highly correlated with them, are known. Thus, as more causal variants are identified, the estimates of disease risk will tend to increase, and the effectiveness with which estimates of risk can be combined (for each of several variants at one locus or across more than one locus) will also increase.

Box 1 | Tantalizing glimpses of new biology

Much remains to be understood about the causal mechanisms underlying findings from GWA studies, but these findings have already indicated pathways that might be involved in disease susceptibility. In addition, intriguingly, it turns out that diseases that differ markedly can share susceptibility loci.

Several examples of new insights into disease-susceptibility pathways are described here. First, in breast cancer, the well-known genes *BRCA1* (breast cancer 1) and *BRCA2* harbour variants with large effects on disease susceptibility, and these have a role in DNA-repair pathways, as do several of the genes that were recently discovered to harbour variants with moderate effects. By contrast, several of the breast-cancer variants discovered in GWA studies might lead to the activation of growth-promoting genes[9], although this is still speculation because the causal variants have not yet been identified.

Second, many variants conferring susceptibility to type 2 diabetes are in loci containing genes that exert their effects by reducing the function or mass of β-cells (which produce insulin in the pancreas), a finding that helps to address long-standing controversies about the mechanisms involved in the pathogenesis of type 2 diabetes[10,11].

Last, in studies of Crohn's disease, findings from GWA studies point to autophagy (the process by which cells digest portions of their own cytoplasm) and innate immune mechanisms as being important in disease aetiology[1].

There are also interesting, and in some cases provocative, patterns in the overlap of associated loci between diseases. For example, one region of chromosome 8q24 contains several independent loci associated with prostate cancer, including one that also confers the risk of developing colon cancer, as well as a nearby (apparently separate) breast-cancer-associated locus. It is unclear whether these loci all confer susceptibility to cancer through a common mechanism, and it is not known which genes are involved, although the nearest gene is the well-known oncogene *MYC*[12].

In addition, researchers are uncovering an increasing number of loci associated with more than one autoimmune disease (and these are outside the major histocompatibility complex, a key region whose variation has long been known to confer susceptibility to multiple autoimmune diseases). But the patterns of overlap are not simple. For example, at one locus, a certain allele confers the risk of developing both Crohn's disease and type 1 diabetes, whereas at another locus, an allele confers type 1 diabetes risk but protects against Crohn's disease[1].

Arguably, one of the most interesting findings from GWA studies is a 120-kilobase region of chromosome 9 that is associated with coronary artery disease[13–15]. The mechanism by which this region contributes to disease is unknown, but the two genes that map closest to the association signal — *CDKN2A* (which encodes p16; also known as INK4A) and *CDKN2B* (which encodes p15; also known as INK4B) — are involved in regulation of the cell cycle. These genes had not previously been suggested as candidates for susceptibility to cardiovascular disease, but they have long been known to have a role in various cancers. To add to the intrigue, other variants in this region of chromosome 9 are associated with type 2 diabetes[16–18].

Another unexpected overlap involves two loci associated with both type 2 diabetes and prostate cancer. One locus contains the gene *TCF2* (transcription factor 2, also known as *HNF1B*), a variant of which confers risk of developing prostate cancer but provides protection against type 2 diabetes. By contrast, in the other locus (containing *JAZF1*, which encodes a zing-finger protein), the risk alleles for these two diseases seem to be distinct[2,19,20].

Recent progress has been made in mapping the DNA sequences associated with disease but, like the early world maps, many crucial details remain unclear.

It is also unclear how to combine these new genetic risk factors with known environmental risk factors. For example, given his particular genotype at each of several cardiovascular-disease-associated loci, what is the risk that a male of 55 years of age who has smoked for 20 years will develop this type of disease? And how does the efficacy of different treatment regimens vary among individuals with different sets of genetic variants? To answer each type of question will require large and expensive studies: prospective cohorts will need to be studied to understand the interplay between environmental risk factors and genetic risk factors, and clinical trials will need to be carried out (using large amounts of genotyping data) to assess how genetic information should be used when choosing treatments. The sooner such studies begin, the better.

The potential for an improved assessment of disease risk has led to several companies offering 'personal genomics', or consumer genomics, services. These companies assay an individual's genome at the position of 500,000–1,000,000 SNPs and provide estimates of disease risk based on known genetic risk factors. It is crucial that such services are carried out responsibly, reliably and to a high scientific standard. In addition, care needs to be taken to ensure that the customers understand what they will, and will not, get for their money, and the customers must be given help to understand how to interpret the results. My instincts are to welcome, rather than to fear, the public interest being generated by personal genomics. Nevertheless, whether prompted by these developments or by an increased use of genomics in routine medical practice, healthcare professionals urgently need improved education about genetics. In addition, much broader educational challenges, among others, will also occur when genetic data flow beyond the medical domain into legal and social domains, and into studies of non-biomedical and psycho social phenotypes.

It is well understood that information about genetic markers gives researchers only a moderate ability to predict whether an individual will develop a particular disease. But personal genomics can be looked at from another perspective. For a particular disease, most individuals will have inherited some sequence variants that confer risk and some variants that provide protection, and they will therefore have an overall risk around the average. A small proportion of people, however, will have inherited mainly variants that confer risk of developing the disease. Using Crohn's disease as an example, if reasonably conservative assumptions are made about the underestimation of disease risk, and if the simplest multiplicative-risk disease model is assumed, then the top 5% of the UK population in terms of Crohn's-disease-associated loci have a 5–8-fold higher risk than average of developing the disease, whereas the top 1% have a 9–15-fold higher risk. The analogous figures based on the published loci for susceptibility to type 2 diabetes are a 2–3-fold increase for the top 5% and a ~4-fold increase for the top 1% (C. Spencer and E. Hechter, personal communication).

Across 50 diseases, making the simplifying assumption that susceptibility to each disease is independent of susceptibility to every other disease, almost everyone will be in the top 5% of risk for at least one disease, and nearly half of all people will be in the top 1% for at least one disease. So, for example, I will be at particularly high risk of developing some diseases because of common variants that I inherited. At present, it is unclear which diseases these are, but with the advent of personal genomics, I can find out. If I am unaware, I cannot be sure whether any changes I make — such as altering my lifestyle, undergoing screening or receiving other interventions — are warranted because of a specific genetic predisposition to disease. Given that it is possible to reduce the risk of developing certain diseases, I can see the value of knowing about the genetic risks now. It should be noted that family

history will continue to be an important risk factor for disease, although it is unclear to what extent the variants that contribute substantially to familial risk (which are likely to include rarer variants with larger effects on risk than those uncovered by GWA studies) overlap with those that are assessed in personal genomics services.

The road ahead

Some of the mysteries of the human genome are now being unlocked, but many puzzles remain to be solved. The success of GWA studies in identifying sequence variants associated with disease has been striking, with the pace and extent of the findings well beyond that of even the most optimistic predictions from just a few years ago. Moving beyond these initial steps will be hard work, however.

Despite the progress made so far, some disenchantment with GWA studies has arisen recently. It might be that the ease with which these studies yielded associations has resulted in their findings being devalued: human nature particularly values hard-fought victories. The growing power of next-generation sequencing technologies also makes the genotyping platforms that underpin GWA studies seem less cutting edge. And, despite the plethora of findings from GWA studies, it is becoming clear that these results fail to explain much of the heritability of the diseases being studied[8].

But it would be a mistake to write off the findings from GWA studies, or the approach itself, too quickly. Progress towards finding the causal variants will undoubtedly increase the amount of heritability that can be explained. And regardless of their effect on disease risk, associated loci can yield, and are already yielding, important insights into disease aetiology, as well as potential drug targets. It is also naive to think that many of the challenges in interpreting GWA study findings would disappear if complete sequence data were available. As just one example, homing in on the causal variant after a GWA study is difficult because of the correlations between nearby SNPs, which result from linkage disequilibrium. This will be just as difficult with resequencing data, for the same reason. In any event, although sequencing costs are decreasing rapidly, it will be some years before it becomes feasible to collect genome-wide sequence data for the number of samples studied in successful GWA analyses.

Parallel studies aimed at explaining the missing heritability — through investigating the role of rare variants, of variation in the copy number of particular sequences, or of epigenetic effects — are also vital. There is a long and challenging road ahead to improving human health, but the rewards will be great. Although the pace of future discovery is unlikely to match that of the past few years, this might still be seen as the time when the tide turned in understanding the genetics of common human diseases. ■

Peter Donnelly is director of the Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.

1. Barrett, J. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40,** 955–962 (2008).
2. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638–645 (2008).
3. Weedon, M. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40,** 575–583 (2008).
4. Lettre, G. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genet.* **40,** 584–591 (2008).
5. Manolio, T. *et al.* A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
6. McCarthy, M. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
7. Hunt, K. A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genet.* **40,** 395–402 (2008).
8. Maher, B. The case of the missing heritability. *Nature* **456,** 18–21 (2008).
9. Turnbull, C. *et al.* Genetic predisposition to breast cancer: past, present, and future. *Annu. Rev. Genomics Hum. Genet.* **9,** 321–345 (2008).
10. O'Rahilly, S. P. *et al.* β-Cell dysfunction, rather than insulin insensitivity is the primary defect in familial type 2 diabetes. *Lancet* **328,** 360–364 (1986).
11. Florez, J. C. Newly identified loci highlight β cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? *Diabetologia* **51,** 1100–1110 (2008).
12. Stratton, M. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nature Genet.* **40,** 17–21 (2008).
13. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316,** 1488–1491 (2007).
14. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316,** 1491–1493 (2007).
15. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 controls. *Nature* **447,** 661–678 (2007).
16. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316,** 1331–1336 (2007).
17. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316,** 1336–1341 (2007).
18. Scott, L. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
19. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk and one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39,** 977–983 (2007).
20. Thomas, G. *et al.* Multiple loci identified in a genome-wide association scan of prostate cancer. *Nature Genet.* **40,** 310–315 (2008).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (peter.donnelly@well.ox.ac.uk).

NATURE|Vol 456|11 December 2008|doi:10.1038/nature07632

A global network for investigating the genomic epidemiology of malaria

The Malaria Genomic Epidemiology Network*

Large-scale studies of genomic variation could assist efforts to eliminate malaria. But there are scientific, ethical and practical challenges to carrying out such studies in developing countries, where the burden of disease is greatest. The Malaria Genomic Epidemiology Network (MalariaGEN) is now working to overcome these obstacles, using a consortial approach that brings together researchers from 21 countries.

Each year, malaria kills about 1 million children and causes debilitating illness in more than 500 million people[1]. Underlying this massive global health problem is a remarkable biological phenomenon, the co-evolution of three eukaryotic genomes[2–9] (Table 1). The disease is caused by single-celled parasites of the genus *Plasmodium*, which invade, and reproduce in, human erythrocytes. The parasites are then transmitted from one person to another by blood-sucking mosquitoes of the genus *Anopheles*.

The evolutionary 'arms race' between the parasite, its vector and the human host is central to the problem of controlling disease. *Plasmodium* populations are continually evolving to resist antimalarial drugs and have sophisticated genetic mechanisms of evading the human immune system, presenting a major problem for the development of a vaccine against malaria[2,7,10]. *Anopheles* populations are likewise evolving to resist the insecticides that are used to control malaria, but they also have genetic defences against the parasite that might provide clues to new control strategies[11,12]. Malaria has also been a strong force for recent evolutionary selection in the human genome[9,13], and uncovering all of the human genetic factors that confer resistance to malaria would provide clues to the molecular basis of protective immunity that would be invaluable for vaccine developers.

The genetic basis of human resistance to malaria can now be investigated systematically at the level of the whole genome, by using genome-wide association (GWA) analysis. In a typical GWA study, the genotype of thousands of individuals is determined at the positions of half a million or more single nucleotide polymorphisms (SNPs)[14] (see page 728). The ultimate goal of GWA analysis is to uncover all of the DNA sequence variants that affect an individual's risk of disease, without sequencing the whole genome, by using statistical inferences based on common patterns of variation in the genome.

An important question that can be addressed by GWA analysis is why only some children develop severe malaria (that is, life-threatening forms of the disease[15,16]) in communities in which every child is repeatedly infected with *Plasmodium falciparum*, the species of parasite that is responsible for most deaths from malaria. Only a small proportion of *P. falciparum* infections progress to severe malaria, and epidemiological data indicate that about 25% of the risk is determined by human genetic factors[17] (Box 1). A typical study design is to recruit individuals with severe malaria (cases) in a hospital setting and to recruit control individuals essentially randomly from the general population. By comparing the frequency of a set of SNPs in cases and controls, it is possible to estimate the effect of different sequence variants on an individual's risk of developing severe malaria. Because the risk of developing severe malaria is probably determined by many genetic factors and environmental factors operating at different stages of infection, the effect of any one factor might be small, so a large number of individuals must be studied to obtain statistically significant results.

Similar approaches could, in principle, be used to investigate the emergence and molecular basis of drug resistance in *Plasmodium* populations or insecticide resistance in *Anopheles* populations. However, this cannot be put into practice until genomic variation in *Plasmodium* and *Anopheles* populations is better understood[3,18–21]. A complicating factor for GWA studies of *P. falciparum* is that in a single infection, the parasites that are transmitted can have different genotypes, so an individual who is infected frequently can carry a parasite population of great genetic complexity. Another is that, in Africa, where malaria is most prevalent, the *P. falciparum* genome has low levels of linkage disequilibrium[19,21]. Linkage disequilibrium is a fundamental concept to consider in GWA analysis; it refers to the correlation between genotypes that is observed at neighbouring positions in the genome. The lower the level of linkage disequilibrium, the more positions in the genome need to be genotyped for an effective GWA study. Recent technological advances in massively parallel sequencing of single DNA molecules[22] might help to overcome both of these problems, by enabling *P. falciparum* to be genotyped at a very large number of positions in the genome and by helping to distinguish the different parasite genotypes that can constitute a single infection.

In this Commentary, we describe how a global research network has been established to investigate the effects of genomic variation in humans on the biology and pathology of malaria. We focus on the human genome because the tools for genotyping and the framework for population genetics are further advanced than those for *Plasmodium* and *Anopheles* species. More specifically, we outline the practical reasons why malaria is more challenging to study by GWA analysis than many other common diseases, and we describe how we have established several projects that bring together large-scale studies carried out in multiple locations to address key scientific questions. We also describe the procedures that we use for standardizing and integrating data from different investigators, as well as the policies that we have developed to deal with issues of sample and data ownership, data release, intellectual property and ethics.

Challenges of GWA studies of malaria

The genetic analysis of human resistance to malaria is challenging at several levels, ranging from the practical and ethical issues of clinical research in the developing world to the statistical genetic issues arising from the great diversity of the populations that are affected.

*A list of participants and affiliations appears in the online version of the paper at www.nature.com/nature.

Table 1 | Malaria involves three eukaryotic genomes

Aspect	*Plasmodium falciparum** (ref. 2)	*Anopheles gambiae*† (ref. 3)	Humans[4,5]
Genomic organization	Haploid (transiently diploid in mosquitoes) 14 chromosomes Plus the mitochondrial genome and apicoplast genome	Diploid 2 autosomal chromosomes, X chromosome, Y chromosome Plus the mitochondrial genome	Diploid 22 autosomal chromosomes, X chromosome, Y chromosome Plus the mitochondrial genome
Genome size (base pairs)	2.3×10^7	2.8×10^8	3.2×10^9
Estimated number of protein-coding genes	5,420	12,945	21,714
Databases of common genomic variants	$\sim 10^5$ unvalidated SNPs listed by dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP)	$\sim 10^6$ unvalidated SNPs listed by dbSNP	$\sim 6 \times 10^6$ validated SNPs listed by dbSNP $\sim 3 \times 10^6$ SNPs investigated by the International HapMap Project[6] (http://www.hapmap.org) $\sim 3 \times 10^4$ structural variants listed by the Database of Genomic Variants (http://projects.tcag.ca/variation)
Examples of evolutionary selection	Resistance to antimalarials[7]: to antifolates, mediated by *DHFR* and *DHPS*; to chloroquine, mediated by *CRT* and *MDR*	Resistance to insecticides[8]: to DDT and pyrethroids, mediated by the *kdr* allele, a variant of a gene encoding a voltage-gated sodium channel	Resistance to severe malaria[9]: mediated by mutations in *HBB* (resulting in different forms of haemoglobin, HbS and HbC); the presence of *HBA1* or *HBA2* (which encode different forms of the α-chain of haemoglobin, resulting in α^+-thalassaemia); by deficiency in *G6PD*

CRT, chloroquine-resistance transporter gene; DDT, dichlorodiphenyltrichloroethane; *DHFR*, dihydrofolate reductase gene; *DHPS*, dihydropteroate synthase gene; *G6PD*, glucose-6-phosphate dehydrogenase gene; HBB, β-globin gene (which encodes the β-chain of haemoglobin); *kdr*, knockdown resistance gene; *MDR*, multidrug-resistance gene; SNP, single nucleotide polymorphism. **Plasmodium* is a genus of protozoan. There are more than 100 species, of which 5 can infect humans: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale* and some forms of *Plasmodium knowlesi*. *P. falciparum* is the most severe form and is responsible for most deaths from malaria. †*Anopheles* is a genus of mosquito. There are about 400 species, of which about 40 transmit the *Plasmodium* parasites that cause human malaria. *Anopheles gambiae* is a major vector of *P. falciparum*.

Recruiting a large number of individuals with severe malaria presents challenges because most of the burden of malaria falls on poor communities with underfunded health services and no systematic medical records. A considerable proportion of children with severe malaria die within hours of reaching a hospital; therefore, for the clinical phenotype of malaria to be classified properly, research information must be gathered at the time of hospital admission. This implies considerable responsibilities on the part of the research team for ensuring standards of medical care, particularly in a resource-poor setting. Also, it is not feasible to take large amounts of blood from children who are ill with malaria, many of whom are anaemic, so it is often necessary to use whole-genome amplification to obtain enough DNA for genotyping at numerous SNP positions. This can reduce genotyping efficiency and thus diminish statistical power, making an even larger sample size necessary[23].

In addition, designing an appropriate 'SNP genotyping' strategy for GWA studies of malaria is complicated by the large amount of genomic variation in Africa. Because of the low levels of linkage disequilibrium in populations in Africa, genotypes need to be sequenced at the positions of more SNPs than in studies of European populations. On the basis of the initial data from the International HapMap Project (http://www.hapmap.org), it was estimated that a GWA study of about 1.5 million SNPs in an African population would be approximately equivalent to a study of 0.6 million SNPs in a European population, in terms of the ability to tag a high proportion of common sequence variants[6]. But it is difficult to estimate how many SNPs will be required to tag all common variants until resequencing studies have generated a comprehensive list of common sequence variants in different African populations[24].

Furthermore, the ethnic diversity of African populations presents numerous statistical challenges for GWA studies. Many African communities consist of several ethnic groups, and minor differences in the ethnic composition of the case groups and the control groups can lead to false-positive genetic associations. To exclude such artefacts, studies need to be designed carefully, and statistical genetic methods that correct for population structure need to be applied[25].

Another problem arises from the genetic differences between populations in Africa, as opposed to within a single population. Signals of association are not expected to be constant across GWA studies carried out at different locations in Africa. For example, differences in haplotype structure can result in variable signals of association around a causal variant of a disease, particularly in genomic regions that have recently undergone evolutionary selection. Also, different populations can harbour different factors that confer resistance to malaria. One example of this is two resistance-associated forms of haemoglobin (haemoglobin S and haemoglobin C) that result from different SNPs at adjacent locations in *HBB*, the gene that encodes the β-chain of haemoglobin — these SNPs have different patterns of distribution in West Africa[26,27].

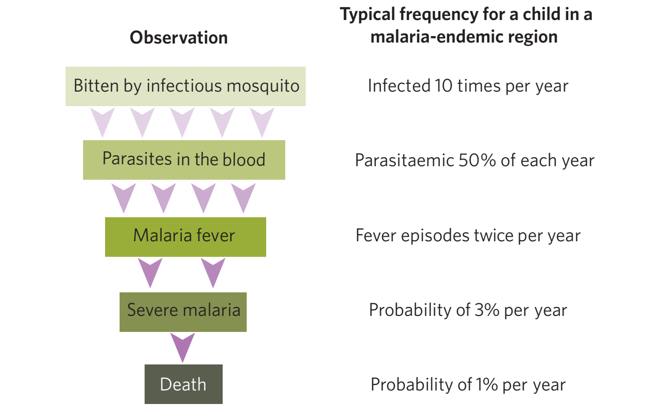
But such differences between populations can be also highly informative. For example, they can aid in uncovering genetic factors that have evolved in specific populations and in investigating interactions between genes and the environment. Importantly, differences in the patterns of linkage disequilibrium between populations can help to distinguish a causal variant from neighbouring polymorphisms. This is necessary because many SNPs that have been associated with particular diseases are not the causal variants but show an association signal simply as a result of correlation with the causal variant (because of linkage disequilibrium). Thus, GWA studies carried out at multiple sites in Africa could provide a rich resource for identifying causal variants.

Developing a global research network

In the past, research into the human genetic factors that affect resistance to malaria has been characterized by multiple research groups each pursuing relatively small studies on their own samples. But the chance of making a discovery, and replicating the finding, is greatly increased if there are effective mechanisms for different research groups to share data and thereby enlarge the number of samples that are studied. The concept of forming a network for sharing data on the genomic epidemiology of malaria — which was to become the Malaria Genomic Epidemiology Network (MalariaGEN) — originated from work that was funded in 2003 by the Bill & Melinda Gates Foundation and by the UK Medical Research Council. The purpose of this funding was to develop web-based software that would allow the integration of clinical and genetic data collected by different research groups. This funding also supported a workshop on the ethical and ownership issues involved in sharing data, which was held in Accra, Ghana, in January 2004 and attended by scientists and clinical researchers from ten research groups in Africa.

MalariaGEN was established in 2005, with joint funding from the Bill & Melinda Gates Foundation (through the Foundation for the National Institutes of Health) and the Wellcome Trust, as part of the Grand Challenges in Global Health initiative[28] (http://www.gcgh.org). The purpose of this joint funding was to discover mechanisms of protective immunity

Box 1 | Progression of malarial disease in a malaria-endemic region



In a malaria-endemic region, there is a large variation in the clinical severity of infections with *Plasmodium falciparum*. When a young child becomes infected, he or she usually becomes ill with fever but eventually recovers. A small proportion of infections progress to severe malaria; that is, forms of the disease that have life-threatening complications, such as profound anaemia or cerebral malaria[15,16]. After repeated infections, older children and adults acquire clinical immunity to malaria, meaning that they can tolerate infection without developing symptoms.

The figure illustrates the likelihood of progression from infection with *P. falciparum* to death for a young child living in a malaria-endemic region, showing frequencies that are representative for such a child. In any given situation, the frequencies will depend greatly on environmental factors, such as mosquito biting rates. Despite the importance of environment, human genetic factors are estimated to account for approximately 25% of the variation among African children in the risk of developing severe malaria[17]. Known genetic factors that confer resistance to malaria, such as the allele that encodes the HbS form of haemoglobin (ref. 9) (Table 1), account for only a small proportion of this variation, implying that many genetic factors involved in resistance remain to be uncovered[17].

Such genetic factors might operate at any stage of the progression from receiving a bite from a malaria-carrying mosquito to dying as a result of severe malaria. For example, individuals who carry the HbS-encoding allele have a tenfold lower risk of severe malaria than those who do not carry this allele, and the mechanism of protection seems to be that HbS acts to suppress the number of parasites in the blood.

to malaria by combining analysis of human genome variation with large-scale epidemiological studies in malaria-endemic regions. Five objectives necessary for achieving this goal were identified: building a global network for sharing data on the genomic epidemiology of malaria; collecting DNA and clinical data from individuals with different phenotypes of malaria; characterizing genetic variation in populations in malaria-endemic regions; identifying genetic variants that provide protection against severe malaria; and defining the immunological mechanisms by which such genetic variants exert their protective effect.

The group of researchers who came together to tackle these objectives, the MalariaGEN investigators, are mainly leaders of clinical, epidemiological or immunological research projects in malaria-endemic areas, and they contribute samples and data to the MalariaGEN programme. Other MalariaGEN investigators contribute expertise and technical resources related to high-throughput analysis of genomic variation, statistical genetics or biomedical ethics. The host institutions of MalariaGEN investigators, the MalariaGEN partner institutions, are located in 15 malaria-endemic countries and 6 other countries (for additional information, see http://www.malariagen.net/resource/1), and the institutions in malaria-endemic countries have well-established study sites, where individuals are recruited to participate in research. Most of these study sites are in sub-Saharan Africa: in Burkina Faso, Cameroon, Gambia, Ghana, Kenya, Malawi, Mali, Nigeria, Senegal, Sudan and Tanzania. There are also MalariaGEN study sites in Papua New Guinea, Sri Lanka, Thailand and Vietnam.

To address the complexities involved in setting up such a global research network, MalariaGEN investigators agreed, at an inaugural meeting in Oxford, United Kingdom, in July 2005, to establish the network in four stages. The first stage was to establish a set of principles and processes, agreed by all investigators, to regulate a central resource of DNA samples and phenotypic data (Box 2 and see http://www.malariagen.net/resource/1). More specifically, this involved standardizing scientific definitions and procedures, enabling partners to gain secure access to the data resource via the Internet, and developing rules about data sharing, intellectual property and appropriate consent.

The second stage was to define a core scientific programme of large-scale experiments and statistical analysis, which would use data and expertise from multiple investigators, and the results of which would belong jointly to all of the investigators involved. Projects that are part of this core programme are called Consortial Projects (Box 2). There are four such projects so far, and each has a specific objective and a plan of action (Table 2). After a Consortial Project has been defined, each investigator decides whether he or she wishes to contribute to the project.

The third stage was to find ways of assisting investigators in malaria-endemic countries to develop clinical and epidemiological studies that would advance the core scientific programme. Investigators were invited to submit funding proposals for projects at their study sites that would contribute to Consortial Projects, using the research infrastructure of the local partner institution and founded on the scientific interests and expertise of the local investigators. Funding was allocated after proposals had been reviewed by a group of investigators that represented the network as a whole (with members from Cameroon, Gambia, Ghana, Italy, Kenya, Malawi, Mali, Sri Lanka, Sudan, Tanzania and the United Kingdom). This group evaluated both the scientific design and the feasibility of the clinical and epidemiological studies proposed, taking into account the infrastructure and expertise of the local partner institution and study site.

The fourth stage was to strengthen the capacity to manage data, and to carry out statistical and genetic analyses, at partner institutions in malaria-endemic countries. A fellowship programme in data analysis was established. After an open application process, a data fellow was appointed at each partner institution. Most of the MalariaGEN data fellows work on the team of a MalariaGEN investigator and have responsibilities for managing the team's data. All data fellows receive training and support in data management, statistical genetics and computing skills. This training is provided by a team of expert statisticians, geneticists and computer programmers who work at the MalariaGEN Resource Centre, which is based at two locations in the United Kingdom, at the University of Oxford and at the Wellcome Trust Sanger Institute near Cambridge. Members of the MalariaGEN resource centre organize regular data-analysis workshops, both in the United Kingdom and at partner institutions in malaria-endemic countries. These workshops provide structured teaching, together with an opportunity for data fellows to share their experiences and to analyse their own data with hands-on assistance from an expert.

Dealing with data

Sharing data is a simple concept but, when many investigators and partner institutions are involved, it can be complex to put into practice. There is the technical issue of how to amalgamate data from different research groups. There needs to be transparency about the ownership and permitted uses of the data and samples contributed by investigators. Procedures need to be established for releasing data and, where appropriate, for protecting intellectual property. This section outlines how MalariaGEN has dealt with each of these areas.

Standardizing and integrating data

Standardizing and integrating data from multiple study sites is central to MalariaGEN's mission. As an example, Consortial Project 1 (Table 2), which is the core project of MalariaGEN's programme, depends on there being a standardized clinical definition of severe malaria. Severe malaria consists of several overlapping clinical syndromes, often referred to as

subphenotypes: these include cerebral malaria (which is characterized by coma), profound anaemia and respiratory distress. Some genetic factors confer resistance generally to severe malaria, whereas others might be specific for a subphenotype. The clinical definition of severe malaria therefore depends on a combination of observations, some of which (for example, respiratory distress) can be quantified less precisely than others (for example, anaemia, through measuring haemoglobin concentration), and there is ongoing research into how to minimize the diagnostic error rate. After consulting MalariaGEN investigators — and after a joint meeting with the Severe Malaria in African Children network[16], in Yaoundé, Cameroon, in November 2005 — a standardized case report form was agreed (see http://www.malariagen.net/resource/1). This form is not intended to replace the case report forms used by individual investigators but rather to provide a template for extracting core information from different clinical data sets in a standardized manner, while giving investigators the freedom to collect data in the way that is most appropriate to their own research.

In a large research network, there will be site-to-site variation in the way in which clinical and epidemiological information is recorded and stored at the local level, so investigators and data fellows are encouraged to have an active role in data standardization and integration. This is facilitated by web-based software developed specifically for this purpose by the MalariaGEN resource centre. Investigators collect data using the database format that is best supported at their institution, and they periodically upload their data via a secure, password-protected interface to a personalized section of the MalariaGEN website, which cannot be accessed by others. Tools are provided for checking data integrity and for transferring data into the database for the relevant Consortial Project. The process of data transfer generally requires the investigator to recode or transform certain variables in their own data set to match the format of the project database, and the web-based software assists and documents this process.

MalariaGEN investigators are also working on the standardization of immunological assays as part of Consortial Project 2, which involves investigating the genetic determinants of the immune response in different populations and environmental settings (Table 2). In the first phase of this project, antibody measurements are being carried out at a central reference laboratory to ensure that data from different study sites can be directly compared. In the long term, the project seeks to develop robust methods and standardized reagents that will enable reference laboratories to be established at partner institutions.

Sharing data and establishing rules of ownership

MalariaGEN is a data-sharing community in which independent investigators with different projects and research objectives contribute to a central repository of DNA samples and a central database of core phenotypic data for each Consortial Project. General principles of data sharing and ownership were agreed at the inaugural meeting of MalariaGEN (Box 2 and see http://www.malariagen.net/resource/1). The major findings of each Consortial Project will be published in scientific journals, with all investigators who contributed to the project listed as authors. In addition, investigators are encouraged to analyse the data that have been generated from their own samples, and to incorporate any additional clinical or experimental data that they have for these samples; these analyses are then permitted to be published independently of the findings of the Consortial Project.

One of the most important considerations when building the database for each Consortial Project was protecting the anonymity of research participants. The MalariaGEN database contains no personal identifiers and is not linked to databases at local study sites. However, one of MalariaGEN's key principles is that investigators should be able to analyse data generated from samples that they contributed and to amalgamate these data with locally held phenotypic data. A standard operating procedure was therefore developed to ensure that the local databases held by partner institutions that contain data generated by MalariaGEN are designed and used according to appropriate ethical guidelines (see http://www.malariagen.net/resource/1).

Box 2 | Key elements of MalariaGEN's policy

MalariaGEN investigators have agreed on a set of principles and processes for sharing samples and data. An important step was to define several Consortial Projects, each of which has a specific objective and project plan (Table 2). Investigators can control how their samples and data are used by MalariaGEN by specifying which Consortial Projects they wish to contribute to.

Definitions

- Consortial Project: a project that uses data and expertise from multiple investigators, is carried out with core MalariaGEN funds, and is agreed by the Project Management Committee, the funding bodies and all of the investigators taking part in the project.
- Contributing investigator: an investigator who contributes data, samples or expertise to a Consortial Project.
- Investigator's own analysis: analysis carried out by a contributing investigator on his or her own samples, either using data generated by MalariaGEN or facilitated in another way by MalariaGEN.

Key principles

- The ownership of physical samples and clinical data contributed to Consortial Projects remains with the contributing investigator.
- The contributing investigator can request that DNA samples be returned at any stage after the agreed experiments have taken place and can use the samples for purposes other than those of MalariaGEN.
- Genotyping data generated by a Consortial Project is fully accessible to the investigator who contributed the samples and may be used for the investigator's own analysis.
- The contributing investigator is responsible for ensuring that samples are taken with the participants' informed consent and for gaining local ethical approval, with support from the MalariaGEN ethics team.
- The laboratories that process samples and analyse data are responsible for the safety and integrity of the samples, and for maintaining security and confidentiality of information.
- Authorship of publications by MalariaGEN will reflect the contributions of all who have provided data and expertise, in accordance with normal academic practice.
- Data at the level of the individual from human GWA studies will be made available to the scientific community through an independent data-access committee.
- Intellectual-property protection will be sought only if it will expedite the translation of a scientific discovery into affordable health benefits for the populations that are most in need.
- For further details of MalariaGEN consortial policies, see http://www.malariagen.net/resource/1

Releasing data and protecting data as intellectual property

Because the scientific benefits of GWA studies are cumulative, the value of a single study can be increased substantially if the data for individual subjects are available to the wider scientific community, provided that the identity of these individuals is securely protected[14,29]. MalariaGEN's policy on this topic was developed in consultation with all MalariaGEN investigators and with ethics-review boards at several MalariaGEN partner institutions (see http://www.malariagen.net/resource/1). In broad terms, the data-release policy seeks to permit research that is consistent with the nature of informed consent and the uses of the samples agreed by the relevant ethics-review boards. A key concern that arose from the consultation was to guard against the data being used in a way that might lead to any form of ethnic stigmatization. Another concern was to ensure that the timeline for data release is fair for investigators in malaria-endemic countries who have contributed resources and data to a project, because these investigators generally have less capacity for analysing genetic data than researchers in rich countries. Balancing the benefits of prompt data release with the need to protect the interests of partner institutions, MalariaGEN's current policy is to release GWA data 9 months after contributing investigators have had access to the complete data set. Data are placed in the European Genotype Archive (http://www.ebi.ac.uk/ega) and are then made available on application

Table 2 | MalariaGEN Consortial Projects

Project	Study sites	Project outline
1. Analysis of human genome variants associated with resistance or susceptibility to severe malaria	Burkina Faso, Cameroon, Gambia, Ghana (Navrongo and Kumasi), Kenya, Malawi, Mali, Nigeria, Papua New Guinea, Tanzania and Vietnam	• Study sites have recruited >10,000 cases and >10,000 ethnically matched controls for genetic association studies of severe malaria, with family trios being recruited at four study sites. • GWA studies of >10,000 individuals to be completed in 2009 for three study sites.
2. Analysis of human genetic factors that determine antibody responses to malaria	Burkina Faso, Kenya, Mali, Senegal, Sri Lanka, Sudan and Tanzania	• Standardized assays for antibodies specific for malaria antigens are being developed. • Assays will be used in a range of community-based studies with various epidemiological designs, to test associations between candidate genes and antibody responses.
3. Analysis of human genome variation in populations in which malaria is endemic	Burkina Faso, Cameroon, Gambia, Ghana, Kenya, Mali, Nigeria, Papua New Guinea, Sudan, Tanzania and Vietnam	• A DNA repository is being established to allow detailed studies of human genome variation in malaria-endemic populations. • Each site has recruited ≥90 individuals and has ethical approval for sequence data to be publicly released.
4. Genetic linkage studies of quantitative traits associated with malaria	Ghana, Senegal and Thailand	• A data-sharing collaboration has been set up between investigators working on genetic linkage studies of the intensity of malaria infection and other quantitative traits.

to an independent data-access committee (as described on the MalariaGEN Data Access web page, http://www.malariagen.net/access). As an additional check and balance, a working group is being established to represent partner institutions and ethics-review boards in malaria-endemic countries, and this group will be kept informed about applications for access to data and consulted about any proposed changes to the data-release policy.

It was important for MalariaGEN to develop guidelines on the circumstances in which data should be protected as intellectual property before publication, with careful consideration of arguments for and against patenting discoveries[30]. On the one hand, if a scientific discovery could lead to health benefits, then every effort should be made to make these benefits available to those who need them most, a process that could involve patenting the discovery. On the other hand, there is an argument for releasing data as openly as possible when there are no immediate applications for improving health and when open access to the data could drive innovations that might lead to health benefits. Arguably, for genomic epidemiology data, the prompt release of scientific findings is, in general, the appropriate course of action, but occasionally there might be discoveries that are exceptions to this. MalariaGEN's current policy is that intellectual-property protection should be sought if all three of the following conditions are satisfied: the discovery must be directly relevant to a medical application; it must be probable that the intellectual property will be licensed for development immediately; and the discovery must have been shown to require intellectual-property protection as a stimulus for further development (see http://www.malariagen.net/resource/1). In such cases, intellectual property will be licensed to non-profit organizations if possible. And, if financial benefits arise, then MalariaGEN will seek to ensure that these benefits flow to the communities who participated in the research.

Engaging with ethical issues

A range of ethical and social issues arise in establishing a network to share data between investigators in many countries. Ensuring ethical standards for the conduct of clinical research in developing countries raises many complex issues[31]. And the accumulation of detailed genomic information about individuals is raising new questions for society in general[32]. This combination of ethical and social issues needs to be addressed appropriately[33]. MalariaGEN has therefore established a team with expertise in medical ethics, which works with investigators and partner institutions to assess the ethical and social issues at different study sites, with the aim of establishing best practices for the ethical conduct of research carried out by MalariaGEN. This ethics team also develops training materials for investigators and ethics-review boards and has held workshops in Kenya, Mali, Thailand and Vietnam. To support investigators in tackling specific ethical issues and to gain an understanding of local practices, members of the team have also visited study sites in Cameroon, Gambia, Ghana, Kenya, Malawi, Mali, Papua New Guinea, Senegal and Sudan.

One of the most important aspects of this work is to find effective ways of communicating with research participants[34]. For example, when a very sick child is brought from her village to a busy government hospital, and her parents are asked whether part of the diagnostic blood sample can be used for a research project, it is often difficult to convey the distinction between medical diagnosis and medical research. Terms such as 'research', 'genetics', 'laboratory' and 'database' might be meaningless unless a concise and effective way is found of translating these concepts into the local language (by using examples and metaphors drawn from local experience), without creating anxiety by information overload. After consulting investigators and ethics-review board members, MalariaGEN has developed a template and guidelines for obtaining informed consent from participants in genetic studies of resistance to malaria (see http://www.malariagen.net/resource/1). To understand how guidelines can be put into practice most effectively, the ethics team is also undertaking empirical research on the process of gaining informed consent at different study sites, with the objective of establishing best practice across MalariaGEN study sites, while being sensitive to local culture and practices.

The ethics team is also working to develop models of consultation at the community level that are appropriate for diverse cultural settings. A sensitive issue for many communities is the potential abuse of genetic data relating to ethnicity, which could result in stigmatization. Qualitative research is being carried out to understand the perspectives of communities and other stakeholders on the collection and use of information about ethnicity in genomic epidemiology projects. The aim is to develop guidelines for the publication and release of data about ethnicity that will provide the maximum scientific benefit while safeguarding the interests of participants and their communities.

Many of the ethical and social challenges confronting MalariaGEN stem from the diversity inherent in a large scientific enterprise with partners in rich and poor countries that span multiple disciplines, from clinical research and community-based research to state-of-the-art genomics and bioinformatics. Often, partners need to agree on an appropriate balance between standardization and shared practices on the one hand, and diversity and sensitivity to local circumstances on the other hand. MalariaGEN's procedures for data integration and guidelines for informed consent are examples of this process.

Looking forward

In September 2008, an ambitious plan for the elimination of malaria was announced — the Global Malaria Action Plan (http://www.rbm.who.int/gmap). This plan, which is supported by major international development agencies and governments around the world, seeks to halve the number of malaria cases worldwide by 2010 and to eliminate deaths from malaria almost completely by 2015. But it cannot succeed without effective insecticides and antimalarial drugs. And even if the plan's goals for the next decade are achieved, the chance of controlling and eliminating malaria over the long term will be greatly increased if an effective vaccine becomes available.

The new science of genomic epidemiology could assist these efforts to eliminate malaria, by providing more effective ways of monitoring the emergence of parasite resistance to antimalarial drugs and of mosquito resistance to insecticides, and by providing new leads for malaria vaccine development based on a better understanding of the natural mechanisms of protective immunity.

If genomic epidemiology is to make a contribution in this way, there need to be mechanisms in place to help researchers both in malaria-endemic countries and worldwide to pool their resources. Research groups in malaria-endemic countries need access to the technical expertise and infrastructure for the large-scale analysis of genomic variation. And research groups worldwide need to combine forces to analyse the massive amounts of data being generated by these studies, leading the way for important discoveries to be made. The MalariaGEN community is endeavouring to learn how to build and maintain the relationships, shared values and best practices that underpin this new type of scientific collaboration. ■

1. Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. & Hay, S. I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434,** 214–217 (2005).
2. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
3. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298,** 129–149 (2002).
4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
5. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
6. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
7. Su, X., Hayton, K. & Wellems, T. E. Genetic linkage and association analyses for trait mapping in *Plasmodium falciparum*. *Nature Rev. Genet.* **8,** 497–506 (2007).
8. Hemingway, J., Field, L. & Vontas, J. An overview of insecticide resistance. *Science* **298,** 96–97 (2002).
9. Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77,** 171–190 (2005).
10. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418,** 320–323 (2002).
11. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298,** 179–181 (2002).
12. Riehle, M. M. *et al.* Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. *Science* **312,** 577–579 (2006).
13. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
14. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
15. Marsh, K. *et al.* Indicators of life-threatening malaria in African children. *N. Engl. J. Med.* **332,** 1399–1404 (1995).
16. Taylor, T. *et al.* Standardized data collection for multi-center clinical studies of severe malaria in African children: establishing the SMAC network. *Trans. R. Soc. Trop. Med. Hyg.* **100,** 615–622 (2006).
17. Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K. & Williams, T. N. Heritability of malaria in Africa. *PLoS Med.* **2,** e340 (2005).
18. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nature Genet.* **39,** 126–130 (2007).
19. Mu, J. *et al.* Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol.* **3,** e335 (2005).
20. Jeffares, D. C. *et al.* Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nature Genet.* **39,** 120–125 (2007).
21. Volkman, S. K. *et al.* A genome-wide map of diversity in *Plasmodium falciparum*. *Nature Genet.* **39,** 113–119 (2007).
22. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5,** 183–188 (2008).
23. Teo, Y. Y. *et al.* Whole genome-amplified DNA: insights and imputation. *Nature Methods* **5,** 279–280 (2008).
24. Bhangale, T. R., Rieder, M. J. & Nickerson, D. A. Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Genet.* **40,** 841–843 (2008).
25. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
26. Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96,** 2358–2363 (2000).
27. Modiano, D. *et al.* Haemoglobin C protects against clinical *Plasmodium falciparum* malaria. *Nature* **414,** 305–308 (2001).
28. Varmus, H. *et al.* Grand challenges in global health. *Science* **302,** 398–399 (2003).
29. Manolio, T. A. *et al.* New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genet.* **39,** 1045–1051 (2007).
30. Chokshi, D. A., Parker, M. & Kwiatkowski, D. P. Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bull. World Health Organ.* **84,** 382–387 (2006).
31. Nuffield Council on Bioethics. *The Ethics of Research Related to Healthcare in Developing Countries.* <http://www.nuffieldbioethics.org/go/ourwork/developingcountries/publication_309.html> (Nuffield Council on Bioethics, 2002).
32. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nature Rev. Genet.* **9,** 406–411 (2008).
33. Chokshi, D. & Kwiatkowski, D. Ethical challenges of genomic epidemiology in developing countries. *Genomics Soc. Policy* **1,** 1–15 (2005).
34. Chokshi, D. A. *et al.* Valid consent for genomic epidemiology in developing countries. *PLoS Med.* **4,** e95 (2007).

Acknowledgements MalariaGEN's primary funding is from the Wellcome Trust (grant number 077383/Z/05/Z) and from the Bill & Melinda Gates Foundation, through the Foundation for the National Institutes of Health (grant number 566) as part of the Grand Challenges in Global Health initiative. Initial work on the web-based software was funded by the Bill & Melinda Gates Foundation (grant number 29015) and the UK Medical Research Council (grant number G0200454). The Wellcome Trust (Sanger Institute core funding) and the Medical Research Council (grant number G0600230) provide additional support for genotyping, bioinformatics and analysis. We thank H. Pearson and colleagues at Bird & Bird for pro bono advice on intellectual property. The MalariaGEN Resource Centre is part of the European Union Network of Excellence on the Biology and Pathology of Malaria Parasites. Individuals who helped to establish MalariaGEN are acknowledged online (see http://www.malariagen.net/resource/1).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to Dominic Kwiatkowski (dominic@sanger.ac.uk).

doi:10.1038/nature07632

The Malaria Genomic Epidemiology Network

Lead Investigators Eric Akum Achidi[1], Tsiri Agbenyega[2], Stephen Allen[3,4], Olukemi Amodu[5], Kalifa Bojang[6], David Conway[6], Patrick Corran[7], Panos Deloukas[8], Abdoulaye Djimde[9], Amagana Dolo[9], Ogobara Doumbo[9], Chris Drakeley[10,11], Patrick Duffy[12,13], Sarah Dunstan[14], Jennifer Evans[2,15], Jeremy Farrar[14], Deepika Fernando[16], Tran Tinh Hien[14], Rolf Horstmann[15], Muntaser Ibrahim[17], Nadira Karunaweera[16], Gilbert Kokwaro[18], Kojo Koram[19], Dominic Kwiatkowski[8,20], Martha Lemnge[21], Julie Makani[22], Kevin Marsh[18], Pascal Michon[3], David Modiano[23], Malcolm E. Molyneux[24], Ivo Mueller[3], Theonest Mutabingwa[12], Michael Parker[25], Norbert Peshu[18], Chris Plowe[26,27], Odile Puijalon[28], Jiannis Ragoussis[20], John Reeder[3], Hugh Reyburn[10,11], Eleanor Riley[10], Jane Rogers[8], Anavaj Sakuntabhai[28], Pratap Singhasivanon[29], Sodiomon Sirima[30], Giorgio Sirugo[6], Adama Tall[31], Terrie Taylor[26,32], Mahamadou Thera[9], Marita Troye-Blomberg[33], Tom Williams[18] & Michael Wilson[19]

Data Fellows Lucas Amenga-Etego[19,34], Tobias O. Apinjoh[1], Edith Bougouma[30], Rajika Dewasurendra[16], Mahamadou Diakite[9], Anthony Enimil[2], Ayman Hussein[17], Deus Ishengoma[21], Muminatou Jallow[6], Enmoore Lin[3], Alioune Ly[31], Valentina D. Mangano[20,23], Alphaxard Manjurano[10,11], Laurens Manning[3], Carolyne M. Ndila[18], Vysaul Nyirongo[24], Tom Oluoch[18], Nguyen T. N. Quyen[14], Prapat Suriyaphol[35] & Ousman Toure[9]

Resource Centre Kirk A. Rockett (Lab Projects Lead)[20], Aaron Vanderwal (Informatics Lead)[20], Taane Clark (Statistics Lead)[8,20], Michael Parker(Ethics Lead)[20,25], Rebecca Wrigley (Network Development Lead)[20], Dominic Kwiatkowski[8,20] (Director), Daniel Alcock[8], Sarah Auburn[8], David Barnwell[20], Susan Bull[20,25], Susana Campino[8], Jantina deVries[20,25], Abier Elzein[17,20], Julie Evans[20], Kathryn Fitzpatrick[20], Anita Ghansah[19,20], Angie Green[20], Lee Hart[20], Eliza Hilton[20], Christina Hubbart[20], Catherine Hughes[20], Anna E. Jeffreys[20], Katja Kivinen[8], Bronwyn MacInnis[8], Magnus Manske[8], Gareth Maslen[8], Marilyn McCreight[20], Alieu Mendy[20], Catherine Moyes[20], Aceme Nyika[8], Claire Potter[20], Paul Risley[7], Kate Rowlands[20], Miguel SanJoaquin[20,24], Kerrin Small[20], Elilan Somaskantharajah[8], Marryat Stevens[20], YikYing Teo[20] & Renee Watson[20]

Project Management Committee Tsiri Agbenyega[2], Dan Carucci[36], Katharine Cook[37], Alan Doyle[37], Ogobara Duombo[9], Jeremy Farrar[14], Michael Gottlieb[36], Kevin Marsh[18], Odile Puijalon[28], Terrie Taylor[26,32] & Dominic Kwiatkowski (Chair)[8,20]

[1]The University of Buea, PO Box 63, Buea, South West Province, Cameroon. [2]Kwame Nkrumah University of Science and Technology, Private Mail Bag, Kumasi, Ghana. [3]Papua New Guinea Institute of Medical Research, PO Box 378, Madang, Papua New Guinea. [4]Swansea Medical School, Swansea University, Singleton Park, Swansea, West Glamorgan SA2 8PP, UK. [5]Institute of Child Health, College of Medicine, University of Ibadan, Ibadan, Nigeria. [6]MRC Laboratories, Atlantic Road, Fajara, PO Box 273, Banjul, Gambia. [7]National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3QG, UK. [8]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [9]The Malaria Research & Training Centre, University of Bamako, PO Box 1805, Bamako, Mali. [10]London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [11]Joint Malaria Programme, Kilimanjaro Christian Medical Centre, PO Box 3010, Moshi, Tanzania. [12]Genome Science Center, Sokoine University of Agriculture, PO Box 3000, Chuo Kikuu, Morogoro, Tanzania. [13]Seattle Biomedical Research Institute, 307 Westlake Avenue North, Seattle, Washington 98109, USA. [14]Oxford University Clinical Research Unit, The Hospital for Tropical Diseases, 190 Ben Ham Tu, Quan 5, Ho Chi Minh City, Vietnam. [15]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Postfach 30 41 2, D-20324 Hamburg, Germany. [16]Faculty of Medicine, University of Colombo, PO Box 271, Kynsey Road, Colombo 8, Sri Lanka. [17]Institute of Endemic Disease, University of Khartoum, Medical Service Science Campus, PO Box 102, Khartoum, Sudan. [18]Kenya Medical Research Institute (KEMRI)–Wellcome Trust Programme, PO Box 230, Kilifi, Kenya. [19]Noguchi Memorial Institute for Medical Research, University of Ghana, PO Box LG 581, Accra, Ghana. [20]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [21]National Institute for Medical Research, PO Box 9653, Dar es Salaam, Tanzania. [22]Muhimbili University of Health and Allied Sciences, PO Box 65001, Dar es Salaam, Tanzania. [23]University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy. [24]Malawi–Liverpool–Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, PO Box 30096, Chichiri, Blantyre 3, Malawi. [25]The Ethox Centre, Department of Public Health and Primary Health Care, University of Oxford, Badenoch Building, Old Road Campus, Headington, Oxford OX3 7LF, UK. [26]Blantyre Malaria Project, PO Box 32256, Chichiri, Blantyre 3, Malawi. [27]University of Maryland School of Medicine, 655 West Baltimore Street, Baltimore, Maryland 21201, USA. [28]Institut Pasteur, Unité d'Immunologie Moléculaire des Parasites, 28 Rue du Dr Roux, 75724 Paris Cedex 15, France. [29]Faculty of Tropical Medicine, Mahidol University, 420/6 Ratchawithi Road, Ratchathewi, Bangkok 10400, Thailand. [30]Centre National de Recherche et Formation sur le Paludisme, Avenue de l'Oubritenga, BP 2208, Ouagadougou 01, Burkina Faso. [31]Institut Pasteur de Dakar, BP 220 Dakar, Senegal. [32]Michigan State University, Department of Internal Medicine, College of Osteopathic Medicine, East Lansing, Michigan 48825, USA. [33]The Wenner-Gren Institute, Stockholm University, SE-106 91 Stockholm, Sweden. [34]Navrongo Health Research Centre, PO Box 114, Navrongo, Ghana. [35]Faculty of Medicine, Siriraj Hospital, Mahidol University, 2 Prannok road, Siriraj, Bangkoknoi, Bangkok 10700, Thailand. [36]Foundation for the National Institutes of Health, 9650 Rockville Pike, Bethesda, Maryland 20814, USA. [37]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.

 NATURE|Vol 456|11 December 2008|doi:10.1038/nature07633

Reverse engineering the genotype–phenotype map with natural genetic variation

Matthew V. Rockman[1]

The genetic variation that occurs naturally in a population is a powerful resource for studying how genotype affects phenotype. Each allele is a perturbation of the biological system, and genetic crosses, through the processes of recombination and segregation, randomize the distribution of these alleles among the progeny of a cross. The randomized genetic perturbations affect traits directly and indirectly, and the similarities and differences between traits in their responses to common perturbations allow inferences about whether variation in a trait is a cause of a phenotype (such as disease) or whether the trait variation is, instead, an effect of that phenotype. It is then possible to use this information about causes and effects to build models of probabilistic 'causal networks'. These networks are beginning to define the outlines of the 'genotype–phenotype map'.

More than a decade into the genomic era, it remains easier to collect genomic data sets than to understand them. The research community has obtained vast quantities of data on genes, transcripts, proteins, metabolites and so on, but has discerned only faint outlines of the networks that connect these factors. Biologists are justifiably enthusiastic about the ability to describe networks of biological molecules that are co-expressed or co-localized, but a central goal of contemporary biology is to connect these observable patterns to form models that predict how biological networks operate as systems. The networks that matter in this context are networks of causal relationships, which can be uncovered by using experiments in which biological systems are perturbed[1–3]. The translation of genotype into phenotype depends solely on these causal relationships; many of the relationships are shaped by the co-expression of genes and physical interactions between cellular components, but many others are determined by intricate networks of cause and effect that are mediated by an organism's physiology, behaviour, and interactions with the environment[4] (Box 1). Inferring causal networks from observations is often called reverse engineering, because the goal is not merely to identify components that are functionally related or situated near to one another but to understand how the system works as an integrated whole.

The classic method for reverse engineering a system is to poke a component with a stick and then to characterize the effect of the perturbation. An alternative is to poke many components simultaneously and at random, repeating the experiment over many random sets of components. Ever since R. A. Fisher put forward his ideas[5] in the 1920s, statisticians have recognized such randomized multifactorial perturbation as the ideal experimental design for uncovering causation. Conveniently, the genetic variation that occurs naturally within a population is a source of multifactorial perturbation[6,7]. The use of natural genetic variation to probe the causal network that links genotype and phenotype has grown recently as large data sets have been generated for many experimental model species, crops and humans[8–10]. In this Review, I discuss recent progress in the application of natural genetic variation to reverse engineer the 'genotype–phenotype map'. After introducing the basic experimental approach, I describe its advantages over traditional genetic screens and show how the resultant data allow tentative inferences to be made about causation. Finally, I discuss the steps that are being taken to gain a mechanistic understanding of the network that connects genotype to phenotype, and I point out potential obstacles to this process, as well as potential shortcuts.

Quantitative genetics of transcript abundance

Genetically characterized populations are the central tool for uncovering the genetic variants that underlie phenotypic variation. A common approach is to cross two inbred lines, each homozygous at every locus, to yield a hybrid that is heterozygous at every locus that differs between the strains. In a typical cross, thousands to millions of genetic loci differ. The ordinary process of meiotic recombination rearranges these polymorphisms within the hybrid germ line, and segments of each of the initial genomes are passed on randomly to the progeny of the hybrids. By tracking the genomic segments with molecular markers, the regions of the genome that contain genetic variants that affect phenotypes (known as quantitative trait loci, QTLs) can be identified[11] (Fig. 1a–d).

An important recent advance in connecting the links between genotype and phenotype has been to measure the 'phenotypic states' of the links, most notably the abundance of the transcripts corresponding to each gene of interest[8–10,12]. Quantitative genetic analysis of genome-wide transcript abundance is sometimes called genetical genomics[6] or expression QTL mapping[9], and the results from this type of analysis — the correlations between genes and transcript phenotypes — can be represented by plotting the physical position in the genome of the gene corresponding to each transcript against the position of the loci associated with variation in transcript abundance (Fig. 1d). In a properly controlled cross, an association between genotype and phenotype implicates genetic variation as the cause of the phenotypic variation; as is the case for all claims based on empirical data, confidence in such a causal inference is defined statistically. The genetic analysis of genome-wide transcript abundance poses distinct technical, computational and analytical challenges that necessitate careful avoidance of potential artefacts[13,14] and that drive innovation in statistical genetics methods[15–19].

Analyses typically show many linkages between the abundances of transcripts and the regions in the genomes where the structural genes for those transcripts reside. Such genes contain QTLs for their

[1]Center for Genomics and Systems Biology, Department of Biology, New York University, 100 Washington Square East, New York, New York 10003, USA.

own transcript abundance. Consequently, the data points align on the diagonal in Fig. 1d. Most of these local linkages will be attributable to *cis*-acting regulatory polymorphisms, although there will be some contribution from polymorphisms that affect the transcript abundance by acting *in trans*[10,20,21].

Data points aligning in vertical bands in Fig. 1d indicate linkage hotspots: that is, regions of the genome at which variation alters the abundance of a large number of transcripts. The loci responsible for these large phenotypic effects might be highly influential regulatory loci[22], or they might be loci at which variation has a marked but non-specific effect on transcript abundances. Such pleiotropic alleles might alter cellular homeostasis in such a way as to shift the steady state for a large number of traits, without this shift being a regulatory effect[10,23].

The quantitative genetics approach described earlier, using natural genetic variation as a source of perturbations, has striking advantages over the classic (one gene at a time) approach[7]. First, the quantitative genetics approach involves massive hidden replication. The effect of each of the alleles present in the initial cross is measured repeatedly because each allele is present in a large number of the phenotyped progeny. A small phenotyping panel of 100 individuals represents on average a 50-fold replication in studying the effect of every allele; similar amounts of replication are not feasible for a one-at-a-time approach.

Second, the presence of simultaneous variation at multiple loci allows the interactions between perturbations (that is, genetic variants) to be uncovered. In the context of gene-expression genetics, such interactions are likely to be common[24,25], and observations have shown that the inheritance of many transcript-abundance traits involves interactions between genes[16,26,27]. A prime example of this gene-interaction phenomenon is genetic redundancy; only by varying the redundant loci simultaneously can their effects be detected.

Last, the simultaneous perturbation of a large number of factors results in the exploration of a larger 'space' of variation than when carrying out single perturbations. Genetically complex traits, which are shaped by variation at many loci, often show transgressive segregation: that is, the random assignment of alleles to the progeny of hybrids results in individuals with unusual collections of alleles, which yield extreme phenotypes. Transgressive segregation characterizes the majority of the transcript-abundance traits in crosses in which its frequency has been examined[26,28]. The large space of phenotypes covered by genetically segregating populations increases the ability to detect relationships between traits. Transcript-abundance data from such populations, for example, are unusually successful at predicting functional relationships between genes[29–33].

Causal ordering

A QTL can affect some traits directly and can affect others indirectly through the effects of intermediate traits. Of particular interest is whether variation in an organismal phenotype, such as a disease state or a behaviour, is an effect of transcript-abundance variation or a cause[34]. Only if the transcript-abundance trait is a cause (not an effect) can it be a target for perturbations that shape variation in the organismal trait, whether in the laboratory, the clinic or an evolving population.

Under a broad set of assumptions, causality shapes correlations in a recognizable way, and its signature is conditional independence. This can be shown by considering three causally related traits: A, B and C. Under standard Markov assumptions, if variation in A causes variation in B, which in turn causes variation in C (this can be written as $A \rightarrow B \rightarrow C$), then when the distribution of B is known, A provides no additional information about C. A and C are independent conditional on B. This statement of conditional independence would not hold if the causal ordering were $B \rightarrow C \rightarrow A$, for example. Conditional independence is by itself insufficient to order causal links uniquely: $A \leftarrow B \rightarrow C$ yields the same conditional independence statement as $A \rightarrow B \rightarrow C$. Nevertheless, conditional independence is a powerful tool for distinguishing direct links between traits from indirect links, even when causal ordering is not possible[1,35,36].

The key advantage of studying genetic perturbations is that many causal orderings are prohibited by the central dogma that genotypic variation can cause phenotypic variation, but, at least within an individual, phenotype does not feed back to affect genotype. Therefore, in a properly controlled cross, genetic perturbations are causally upstream of phenotypes and provide a terra firma into which causal networks can be rooted[34,37–40] (Box 2).

There are several methods for causally ordering pairs of phenotypes measured in segregating populations, including approaches that simultaneously map QTLs and fit causal models[18], approaches that apply formal statistical tests to identify direct causal links[39], and approaches that fit various causal models to triplets that comprise two traits and a QTL and then compare the fit of the models using information-theory criteria[34].

Analysis of the correlations between multiple traits that share an underlying QTL can also help to identify the causal gene within a QTL interval[33,37,41–44] — the major challenge in quantitative genetics today. Many such analyses incorporate sources of information in addition to the correlations, including data on the binding sites of transcription factors,

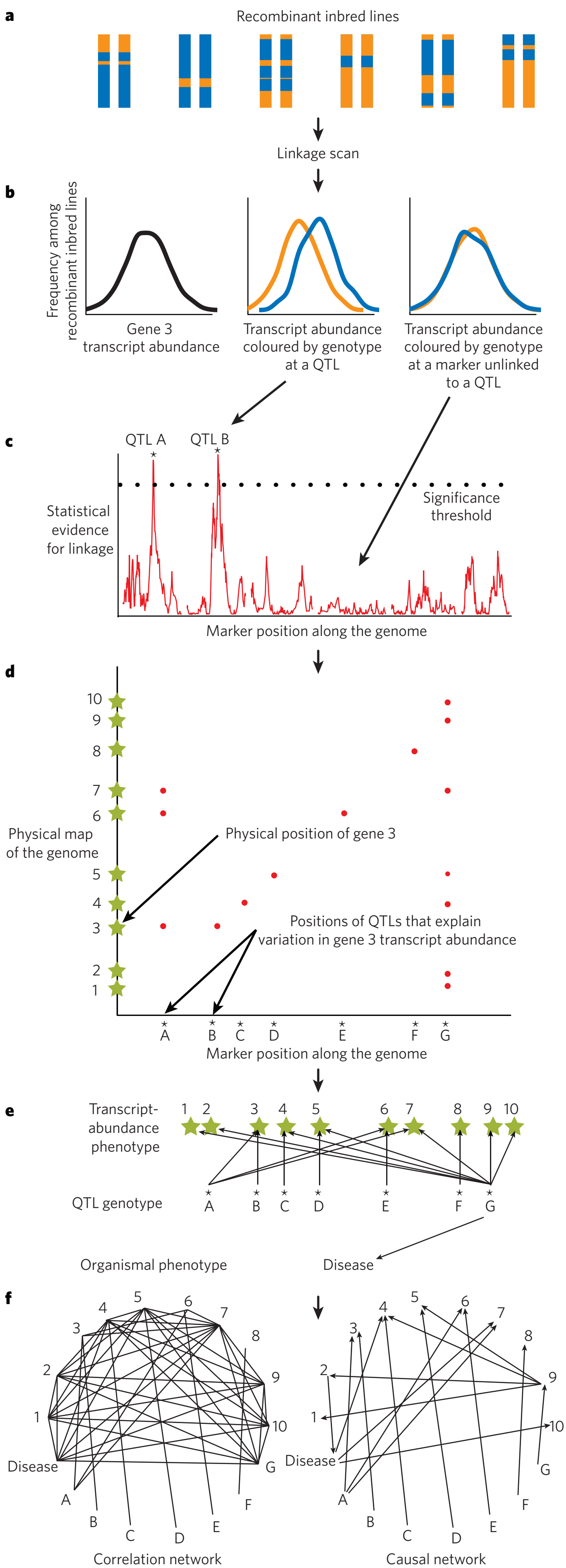
Box 1 | Do causal networks exist?

The concepts of causality and networks are controversial. Many biologists are sceptical about inferring causality from statistics and about how general and useful network models might be, so it is valuable to consider how causality and networks can be interpreted in the context of the relationship between genotype and phenotype.

When carrying out a biological experiment, most people are content with the everyday theory of causation that dictates that one fact precedes a second and alters its probability (setting aside questions about the ontological status of probabilities). Empirical claims about links between causes and effects rely on assumptions (often implicit) and on statistical measures of confidence. Even when testing a simple single-gene perturbation, such as in a gene-knockout organism, it is assumed that inductive reasoning will lead to knowledge, that genotypes are fixed and that they causally precede phenotypes, and that all variables are fully controlled (for example, by comparing the experimental organisms with wild-type full siblings, by blinding observers to differences in treatment, and by carrying out randomized replications of the entire experiment). A researcher's confidence that the wild-type organism and the mutant organism show real differences is influenced by statistical tests (for example, the P value from a t-test), which are themselves typically laden with assumptions.

In inferring probabilistic causal networks, the assumptions are more numerous, but the conceptual framework is the same. At the end of an analysis, the outcome is a claim about cause and effect, and this claim is accepted to an extent that is defined by the researcher's comfort with the assumptions and the statistics. For genetics experiments, the limits of comfort for most researchers lie not far beyond the single-gene perturbation experiment, and making inferences about a causal network is typically seen as a technique for nominating candidate genes for follow-up experiments.

Whether networks exist is a popular topic at biology department happy hours, but it is not necessary to subscribe to the reality of a Platonic Network, an ideal form independent of the material world, in order to embrace the idea that there is a many-to-many relationship between causes and effects in biology. The more pressing question is which components are needed to represent such a network in a predictive model: molecules, interactions, dynamics, all of these, or more? Conveniently, the set of variables in a causal network is entirely circumscribed by the set of things that vary. A molecule or an event can be required for a biological process, but if it does not vary, then it cannot be a cause. In that sense, causal networks in genetics are analogous to the geneticist's concept of heritability, which describes not the dependence of a trait on inherited genes but the proportion of the trait's variation that can be explained by genetic variation in the observed sample. A causal network that is inferred from a genetically segregating population will therefore depend on the genetic variation that is present in the population and on the distribution of variation in the environment across the sampled individuals. Within that framework, a causal network constitutes a predictive model of the consequences of perturbing the represented variables.



the interactions between proteins, and the presence of polymorphisms in the sequences of each gene in the QTL. As a starting point, a gene that is found at the same location as a QTL for its own abundance is a strong candidate for the causal gene underlying variation in other traits that link to the same location, with the gene's transcript abundance as a candidate causal trait[29].

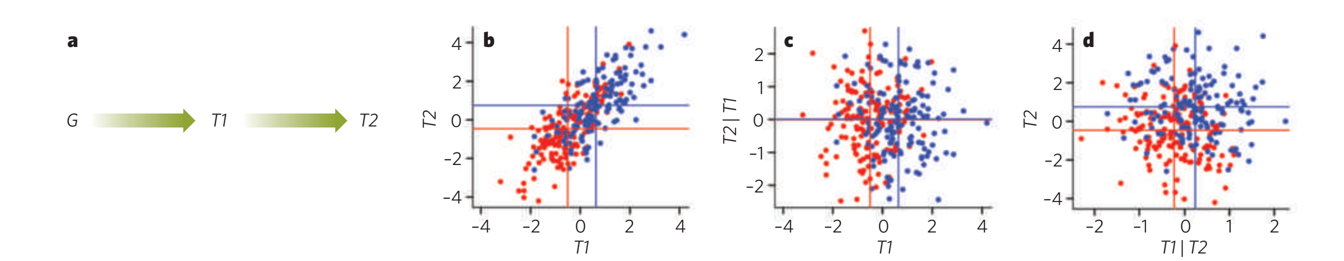
The use of phenotypic correlations to identify causal transcripts is especially promising in the context of genome-wide association studies. These studies have recently uncovered a wealth of high-confidence, replicated associations between genetic variants and diseases in humans (see page 728), but the disease-associated variants are often in non-coding regions with unknown function[45]. The mechanisms that link genotype and disease in these cases can be identified by taking advantage of the structure of the correlations among transcript-abundance traits and disease states in human populations[32,46–48]. In population-based association mapping, however, correlations between genotype and phenotype can arise from external causes that are common to the correlated variables, for example from the stratification of populations by age or ethnicity. Consequently, the causal 'anchor' provided by genotype in these cases is less secure than in the experimental setting of inbred line crosses. But studies in animal models can be used to corroborate findings, providing reassurance[31,49].

Causal networks

To gain a predictive, systems-level understanding of biological causation, researchers need to integrate the entire ensemble of genetic variants and phenotypic traits (and not just to causally order trait pairs). Several approaches aim at this more general goal, and Bayesian networks provide the most popular framework[1]. A Bayesian network is a graph of random variables, each representing a phenotype in this case, that are connected by directed edges. A set of probability distributions describes the state of each variable conditional on the variables with edges leading to it. The graph and probability distributions define a conditional probability statement. One problem with using Bayesian network graphs is that several directed graphs can be described by a

Figure 1 | From genetic randomization to causal network. A genetically randomized population, such as a panel of recombinant inbred lines whose chromosomes carry random segments of genome from two progenitor strains (depicted in orange and blue) (**a**), is a starting point for linkage analysis of a phenotype. In this case, the phenotype is the abundance of transcripts corresponding to a gene denoted as gene 3. The abundance of gene 3 transcripts varies between the recombinant lines (**b**, left). Each point along the genome is tested to see whether it affects the abundance of gene 3 transcripts (**b**, centre and right), and statistical evidence is uncovered for the linkage of gene 3 with two regions (indicated by asterisks) (**c**). These regions are called QTLs. If a similar experiment is carried out for many transcript-abundance phenotypes (not just for gene 3, but for genes 1–10), the positions of the QTLs that affect transcript abundance (asterisks) can be plotted against the physical positions of the gene corresponding to each transcript (green stars) (**d**). In such a plot, the data (red dots) along the diagonal line represent local linkages, typically due to *cis*-acting regulatory polymorphisms. Vertical alignments in the plot indicate linkage hotspots. The plot depicted implies a high-level causal network (shown in **e**), in which QTL variation is the cause of variation in transcript-abundance phenotype. Transcript-abundance QTLs can co-localize with QTLs for organismal phenotypes such as a disease (not shown); for illustrative purposes, disease is shown linked to QTL G. A goal of the reverse engineering of causal networks is to include phenotypes as variables, for example to determine whether the transcript abundances that are affected by QTL G are causes or effects of the disease. Although the traits are densely connected by correlations — as is evident from the hypothetical correlation network that is depicted (**f**), which connects all traits that share perturbations — a causal network (**f**) reveals that QTL G acts directly on gene 9, the transcript abundance of which affects genes 1, 2, 4 and 5. The transcript abundance of gene 2 is a cause of disease, which in turn alters the transcript abundances of genes 4, 7 and 10. Many transcripts are correlated with disease, but only perturbations of genes 2 and 9 will affect disease outcome.

Box 2 | Causal ordering yields conditional independence



The basis for causal inference can be shown graphically. Consider a population of haploid individuals with a single causal locus (*G*) that has two alleles. The allelic state at this locus causes variation in the abundance of the corresponding transcript (*T1*), and additional sources of variation (genetic, environmental and stochastic) also influence *T1*. Data can be simulated with the allelic effect modelled as β_1 and the additional variation modelled as normally distributed noise, ε. Thus, $T1 = \beta_1 G + \varepsilon$, where *G* is in a indicator variable for genotype. Variation in the abundance of *T1* causes variation in a downstream trait, *T2*, which is also affected by other sources of variation, so $T2 = \beta_2 T1 + \gamma$, where γ is an additional noise term. The causal ordering is shown as a scheme in panel **a** of the figure, and simulated data are plotted in panel **b** of the figure.

Data were simulated to represent 300 individuals. Each data point is coloured according to genotype (blue for one allele of *G* and red for the other), and the mean values for each trait are indicated by a coloured line for each genotype. (For this simulation, genotypes were assigned indicator variables: red was assigned −1, and blue was assigned 1. The parameters used were $\beta_1 = 0.5$ and $\varepsilon \sim N(0, 1)$, yielding a zero-mean trait *T1*. For *T2*, $\beta_2 = 1$ and $\gamma \sim N(0, 1)$, so *T2* is simply *T1* with additional noise.)

The causal links mean that the traits, *T1* and *T2*, are correlated with one another and that both are correlated with genotype (figure, **b**). Nevertheless, the relationship between *T2* and *G* is entirely mediated by *T1*. *T2* conditional on *T1* is independent of genotype; the mean phenotype for each genotype is the same (figure, **c**). Conversely, the distribution of *T1* conditional on *T2* remains dependent on genotype (figure, **d**). It is the noise component of *T1* variation, ε, propagated through *T2* that makes this mode of analysis possible, and it is the causal anchor of the genotype that gives it direction.

single conditional probability statement (as discussed earlier), so observations of the random variables (the transcript abundances in this case) cannot uniquely identify the directed network that underlies the graph. Moreover, a second problem is that Bayesian network graphs are acyclic and therefore cannot model feedback regulation; the consequences of this limitation for the utility of Bayesian network are unclear[1,31,37]. A third problem is that the space of possible network graphs is large, making causal-network inference a computationally intractable problem[1]. These difficulties notwithstanding, transcript-abundance measurements from genetically segregating populations are uniquely suited to uncovering directed Bayesian networks for two main reasons[29,37]. First, a trait that is caused by another trait should share an underlying genetic perturbation: a QTL. This simple filter excludes a huge proportion of the space of possible networks, making the problem tractable. Second, genetic perturbations anchor causal networks (as discussed earlier), giving direction to the edges. Although large-scale causal-network inference remains challenging, the incorporation of genetic data clearly improves the quality of the predictions over those derived solely from trait correlations[50].

In parallel with Bayesian network models, structural equation models have been applied to transcript-abundance data from segregating populations[38,51]. These models involve systems of linear equations organized into a network structure; a linear model is fitted with variables that simultaneously function as predictors and responses. Although structural equations, unlike Bayesian networks, have the advantage of allowing feedback cycles to be modelled, they require the standard assumptions of linear modelling. Therefore, when nonlinear causal dynamics underlie transcript abundances, problems can arise. Bayesian networks typically deal with nonlinearity incidentally, by classifying all of the data into simple discrete categories (for example, upregulated, downregulated and unchanged), a simplification that has its own drawbacks[50].

An alternative approach is to generate a simple network from pairwise trait correlations and then to trim this network by testing for conditional dependence relationships[35]. The resultant undirected graph can then be directed by anchoring the edges in QTLs[40]. There are clear computational advantages to starting with a network that is derived from pairwise correlations. Because correlations between genes typically show modularity — with clusters of highly correlated genes being largely uncorrelated with other such clusters — pairwise analyses can break intractably large problems into problems that focus on individual modules[29,52,53].

Two recent studies on disease phenotypes in humans and mice used this shortcut of partitioning the transcript-abundance data into modules of correlated traits[31,32]. Breaking the problem down further, the authors compared causal orderings for pairs of transcript abundances and disease phenotypes, to see whether each transcript could be placed causally upstream of the disease state. A single module, evident in both mouse and human data, was significantly enriched for putatively causal traits; subsequent experimental manipulations corroborated these inferences. Although this is far from a complete reverse engineering of the genotype–phenotype map, these empirical successes (reducing genomic data to the two-trait ordering problem) point to a coming age in which prediction will be a common tool.

Quantitative genetics is evolutionary genetics

The central dogma that genes are causes of phenotypes within an individual aids in the anchoring of directed networks. But among individuals, phenotypes feed back by selection to shape genes. Natural variation therefore samples a biased subset of possible genetic perturbations, a subset that is enriched for those variants that are not strongly deleterious. Under the classic infinitesimal model of the genotype–phenotype map, variation derives from mutations of small effect with limited pleiotropy[54]. If this model holds, the modularity of networks inferred from natural variation[29,31,32,37] might be an epiphenomenon of natural genetic perturbations, the effects of which are less systemic than those of random mutations.

The effect of selection is evident in the numbers and types of QTL detected in typical studies. James Ronald and Joshua Akey found that the proportion of genes showing local QTLs in a yeast cross is smaller than expected under neutrality, implying that negative selection keeps certain perturbations at low frequency[55]. Similarly, genes that are crucial regulators of essential processes are likely to be under-represented among genetically variable genes. Genes that encode transcription factors, for example, are clearly candidates when looking for the genes involved in varying gene expression, but these genes are largely absent from expression QTLs[18,56]. Nevertheless, links between transcription factors and target genes can be detected by causal inference approaches, even when the transcription-factor locus contains no genetic variation. The only requirement is that some phenotypic measure of the transcription factor's activity (such as the abundance of the corresponding transcript) is causally intermediate between the genotype and the target gene's phenotype[31].

The effect of selection on the filtering of genetic perturbations varies according to the type of experimental population studied. Inbred line crosses often involve genetically divergent lines, chosen to maximize phenotypic or genotypic differences. The alleles that contribute to divergence are likely to differ in their allelic effects from those that contribute to standing variation (the ordinary genetic diversity present within a population)[57–59]. Crosses between divergent inbred lines are more likely to uncover rare large-effect mutations (that is, linkage hotspots) than are samples of individuals from large populations[32,55,60]. Such crosses might also be biased towards a subset of perturbations with large effect, not individually but in combinations, as a result of coadaptation[61]. Pervasive genetic interaction means that causal links between pairs of genes will be poorly modelled, although there has been progress towards solving this problem recently[16].

The suite of naturally occurring perturbations is also shaped by population genetics phenomena that are unrelated to the fitness effects of the perturbations themselves. Genetic variants tightly linked to other variants that are the target of selection are evolutionarily coupled to them, yielding a correlation between local recombination rate and levels of genetic variation[62,63], and mutagenic recombination can produce the same pattern[64]. Genes present in regions that undergo recombination at a low rate are less likely to contribute to the pool of genetic perturbations than genes in regions where recombination occurs frequently. If gene location is nonrandom with respect to recombination rate, then there might be fewer perturbations belonging to particular functional classes[61].

Prospects for genetical systems biology

Causal network inference faces many difficulties, and its application to gene expression in segregating populations introduces additional challenges. Despite many published reports of empirical successes, it is important to consider the pitfalls that unpublished studies might have encountered. Many of these pitfalls are now well recognized, and there are clear paths around them, on both the experimental front and the analytical front, towards a richer understanding of the genotype–phenotype map.

One of the most problematic assumptions that is made when drawing causal inferences from gene-expression data is that measurement errors are similarly distributed across traits. If a causal trait is poorly measured and the trait it affects is well measured, then the measurements of the 'effect trait' might report the true values of the causal trait more accurately than measurements made directly on the causal trait itself[34,39,51]. In such cases, conditional correlation approaches can yield inverted inferences (Fig. 2). There are good biological reasons to be concerned about this situation. For example, regulatory molecules are often present at low abundances in a cell, but their effects are amplified by the dynamics of gene regulation. Thus, variation in the number of transcripts encoding a low-abundance transcription factor might be the cause of variation in the number of transcripts encoding a high-abundance structural protein. The difficulty arises in that the high-abundance transcripts might be easy to measure with great precision, whereas the low-abundance transcripts might be present at the threshold of detection. One possible solution is to return to the early designs of microarray experiments, when technical replicates were routine. Taking multiple independent measurements of transcript abundances from a single biological sample would generate empirical parameters for gene-specific error models. The potential to be misled when making causal inferences underscores the point that a causal inference yields an assumption-laden probability statement, and stronger claims about causality need to be experimentally validated[3].

Another concern is that data for transcript-abundance mapping are derived from mixed populations of cells. Consequently, the measurements describe cellular mixtures, the characteristics of which are determined by developmental and cellular demographics[31,32,65]. This is as true for yeast, which has been studied by measuring unsynchronized cultures (which contain cells at different stages of the cell cycle), as it is for animals and plants, in which tissues or whole organisms are assessed. A study of mice that varied genetically in the cell-cycle timing of their haematopoietic stem cells took advantage of the issue of mixed cell populations to identify which genes were expressed differentially in the different cell populations[65].

A related issue is that networks inferred from segregating populations are static representations. Without time-series data, there can be no account of dynamics. Inferred causal links correspond to perturbations that alter the steady state. In a system with feedback, which is likely to include almost all biological steady states, the causal links will describe only the causes that 'win out' over others in shifting the phenotypic balance[31]. For practical considerations of whether it is possible to predict how novel perturbations will affect steady states, these low-dimension projections of the network might be adequate; however, for true reverse engineering, more-complete models are needed. Time-series data have proved exceptionally important for solving the general problem of causal-network inference[2], and it is clear that integrating such information into studies of genetic variation will improve the knowledge gained from these studies.

Many network-inference methods depend on the inclusion of all causal variables in the models[1], but such completeness is not plausible for most biological networks. Some phenotypic causes will be overlooked because they are not measured: for example, unannotated genes, genes present in structural polymorphisms that are absent from reference genomes, and transcripts of small RNAs. In addition, the abundances of metabolites are typically not considered, although much progress has been made in this area recently[40,66]. Regulatory events for which there are no transcriptional indications (for example, post-translational regulation) will also be missed, although again progress is occurring on

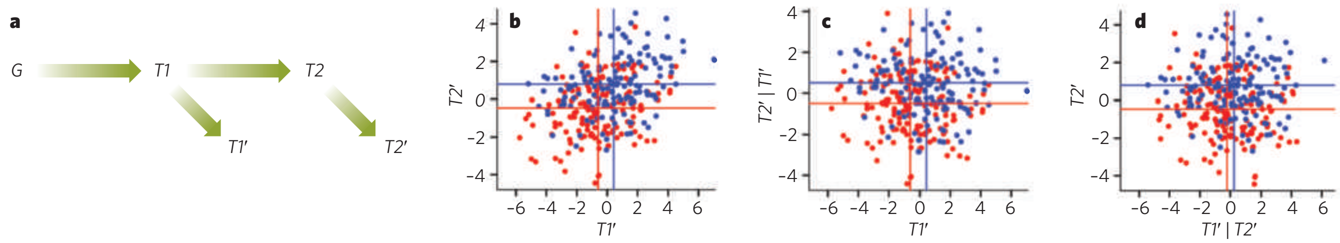


Figure 2 | Measurement error can confuse causal inference. The effect of errors in measurement on causal inference is depicted for the population, parameters and conditions set out in Box 2. In brief, a population of haploid individuals has a single causal locus (*G*) with two alleles, and the allelic state at this locus causes variation in the abundance of the corresponding transcript (*T1*), which subsequently affects the abundance of another transcript (*T2*). **a**, The measured values of *T1* and *T2* in this example were simulated as their true values from Box 2 plus normally distributed error, yielding *T1′* and *T2′*. For *T1′*, the error is normally distributed with a variance of 2, whereas the variance of *T2′* is tenfold lower. The causal ordering of this scheme is shown. **b**, *T1′* and *T2′* are correlated with one another and linked to the genotype, which is represented by the colours (blue for one allele of *G* and red for the other). However, the conditional correlations are now misleading with respect to the true causal network (shown in **a**). **c**, *T2′* remains dependent on genotype after taking *T1′* into account, which is unexpected given the causal ordering (**a**) (and given that *T2* conditional on *T1* does not depend on genotype; Box 2 figure, panel c). **d**, *T1′* is nearly independent of genotype after taking *T2′* into account, which is also unexpected given the causal ordering (**a**) (and given that *T1* has been shown to depend on genotype; Box 2 figure, panel c). In total, the effect of the differences in measurement error is to make *T2′* a better measure of the true *T1* than *T1′* itself.

this front[67,68]. Some missing causal links will arise from an organism's history, because genotypic variables act across an individual's lifespan. For example, the expression of a gene that acts early in an organism's life might show no correlation with the phenotypes it affected at the time of measurement, but the gene–trait relationship will remain[6]. The catalogue of genetic causes will also be incomplete. For large numbers of gene-expression traits, the underlying genetic variants that affect their expression will be undetected because of insufficient power[26]. A simple solution to this problem is to use larger sample sizes, and on this front progress is also being made[69].

Finally, the realm of biological causes is enormous[4], and most experiments limit exploration to a tiny controlled corner of this realm (Box 1). Studies of common lines in multiple environments are now providing the first steps towards integrating genetic perturbations and environmental perturbations into a single view of causal networks[70,71]. Analysing transcript abundances across multiple tissues and sexes is equally important for studying context-dependent causal networks, because each cell type provides a distinct environment for the genome[32,72–74]. Ultimately, as data collection becomes more rapid and less expensive, researchers will be able to study a broader range of conditions.

Natural genetic variation is the stuff of evolution and the cause of heritable susceptibility to diseases. Its properties are fundamentally important to a wide range of biological disciplines. It is therefore fortunate that natural genetic variation has the character of an ideal multifactorial perturbation, providing a natural experimental design that is helping researchers to uncover the mechanistic basis of the map that connects genotype to phenotype. As the molecular catalogues of genomics yield to the integrated models of systems biology, natural genetic variation will have an increasingly central role. ■

1. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7,** 601–620 (2000).
This paper provides a clear overview of Bayesian network formalisms, the main framework for causal network inference at present, and demonstrates how they can be applied to gene-expression data.
2. Bonneau, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131,** 1354–1365 (2007).
3. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18,** 389–401 (2007).
4. Oyama, S. *The Ontogeny of Information: Developmental Systems and Evolution* (Duke Univ. Press, 2000).
5. Fisher, R. A. The arrangement of field experiments. *J. Ministry Agric. Great Britain* **33,** 503–511 (1926).
6. Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* **17,** 388–391 (2001).
This was the first article in which the many advantages of using natural variation to probe gene-expression networks were articulated.
7. Jansen, R. C. Studying complex biological systems using multifactorial perturbation. *Nature Rev. Genet.* **4,** 145–151 (2003).
8. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
9. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302 (2003).
References 8 and 9 provided the first empirical results showing the power of genetic analysis of genome-wide gene expression.
10. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Rev. Genet.* **7,** 862–872 (2006).
11. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, 1998).
12. Stamatoyannopoulos, J. A. The genomics of gene expression. *Genomics* **84,** 449–457 (2004).
13. Perez-Enciso, M. *In silico* study of transcriptome genetic variation in outbred populations. *Genetics* **166,** 547–554 (2004).
14. Alberts, R. *et al.* A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171,** 1437–1439 (2005).
15. Carlborg, O. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21,** 2383–2393 (2005).
16. Storey, J. D., Akey, J. M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **3,** e267 (2005).
17. Kendziorski, C. M. *et al.* Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62,** 19–27 (2006).
18. Kulp, D. C. & Jagalur, M. Causal inference of regulator–target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7,** 125 (2006).
19. Jia, Z. & Xu, S. Mapping quantitative trait loci for expression abundance. *Genetics* **176,** 611–623 (2007).
20. Doss, S., Schadt, E. E., Drake, T. A. & Lusis, A. J. *Cis*-acting expression quantitative trait loci in mice. *Genome Res.* **15,** 681–691 (2005).
21. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* **1,** e25 (2005).
22. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430,** 743–747 (2004).
23. Churchill, G. A. The genetics of gene expression. *Mamm. Genome* **17,** 465 (2006).
24. Omholt, S. W., Plahte, E., Øyehaug, L. & Xiang, K. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* **155,** 969–980 (2000).
25. Gjuvsland, A. B., Hayes, B. J., Omholt, S. W. & Carlborg, O. Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* **175,** 411–420 (2007).
26. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102,** 1572–1577 (2005).
27. Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701–703 (2005).
28. West, M. A. *et al.* Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175,** 1441–1450 (2007).
29. Lum, P. Y. *et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J. Neurochem.* **97** (suppl. 1), 50–62 (2006).
30. Huttenhower, C. *et al.* Nearest neighbor networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* **8,** 250 (2007).
31. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
32. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
References 31 and 32 integrated association mapping in human populations and linkage mapping in mice to identify suites of functionally related genes that are causally implicated in disease.
33. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genet.* **40,** 854–861 (2008).
34. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37,** 710–717 (2005).
35. de la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20,** 3565–3574 (2004).
36. Magwene, P. M. & Kim, J. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.* **5,** R100 (2004).
37. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105,** 363–374 (2004).
This paper was the first to integrate expression QTL data and phenotypic correlation data into causal modelling, as well as to describe the crucial role of genetic perturbations in anchoring causal links in the Bayesian network context.
38. Li, R. *et al.* Structural model analysis of multiple quantitative traits. *PLoS Genet.* **2,** e114 (2006).
39. Chen, L. S., Emmert-Streib, F. & Storey, J. D. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* **8,** R219 (2007).
This paper details a conservative analysis pipeline for uncovering high-confidence causal links with a well-defined false-discovery rate.
40. Ferrara, C. T. *et al.* Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genet.* **4,** e1000034 (2008).
41. Bing, N. & Hoeschele, I. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170,** 533–542 (2005).
42. Li, H. *et al.* Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.* **14,** 1119–1125 (2005).
43. Tu, Z. *et al.* An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics* **22,** e489–e496 (2006).
44. Suthram, S. *et al.* eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.* **4,** 162 (2008).
45. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
46. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39,** 1202–1207 (2007).
47. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39,** 1208–1216 (2007).
48. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature Genet.* **39,** 1217–1224 (2007).
49. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
50. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* **3,** e69 (2007).
51. Liu, B., de la Fuente, A. & Hoeschele, I. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178,** 1763–1776 (2008).
52. Ghazalpour, A. *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2,** e130 (2006).
53. Lee, S. I. *et al.* Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl Acad. Sci. USA* **103,** 14062–14067 (2006).
54. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
55. Ronald, J. & Akey, J. M. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE* **2,** e678 (2007).
This paper is a founding contribution to the field of functional population genomics; it addresses the genomic basis of phenotypic evolution from the perspective of the functional alleles segregating in populations.
56. Yvert, G. *et al.* *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35,** 57–64 (2003).
57. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3,** 11–21 (2002).
58. Ohta, T. Origin of the neutral and nearly neutral theories of evolution. *J. Biosci.* **28,** 371–377 (2003).
59. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genet.* **40,** 346–350 (2008).

60. Schliekelman, P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178,** 2201–2216 (2008).
61. Petkov, P. M. *et al.* Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet.* **1,** e33 (2005).
62. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356,** 519–520 (1992).
63. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
64. Kulathinal, R. J., Bennett, S. M., Fitzpatrick, C. L. & Noor, M. A. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc. Natl Acad. Sci. USA* **105,** 10051–10056 (2008).
65. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genet.* **37,** 225–232 (2005).
66. Wentzell, A. M. *et al.* Linking metabolic QTLs with network and *cis*-eQTLs controlling biosynthetic pathways. *PLoS Genet.* **3,** 1687–1701 (2007).
67. Foss, E. J. *et al.* Genetic basis of proteome variation in yeast. *Nature Genet.* **39,** 1369–1375 (2007).
68. Stylianou, I. M. *et al.* Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification. *Genetics* **178,** 1795–1805 (2008).
69. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36,** 1133–1137 (2004).
70. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2,** e222 (2006).
71. Smith, E. N. & Kruglyak, L. Gene–environment interaction in yeast gene expression. *PLoS Biol.* **6,** e83 (2008).
72. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37,** 243–253 (2005).
73. Cotsapas, C. J. *et al.* Genetic dissection of gene regulation in multiple mouse tissues. *Mamm. Genome* **17,** 490–495 (2006).
74. Wang, S. *et al.* Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2,** e15 (2006).

Acknowledgements I thank the Jane Coffin Childs Memorial Fund for Medical Research and New York University for support, and L. Chen for discussion.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (mrockman@nyu.edu).

FCA does not bind abscisic acid

Arising from: Fawzi A. Razem, Ashraf El-Kereamy, Suzanne R. Abrams & Robert D. Hill *Nature* **439,** 290–294 (2006)

The RNA-binding protein FCA promotes flowering in *Arabidopsis*[1]. Razem *et al.*[2] reported that FCA is also a receptor for the phytohormone abscisic acid (ABA). However, we find that FCA does not bind ABA, suggesting that the quality of the proteins assayed and the sensitivity of the ABA-binding assay have led Razem *et al.* to erroneous conclusions. Because similar assays have been used to characterize other ABA receptors[3,4], our results indicate that the ABA-binding properties of these proteins should be carefully re-evaluated and that alternative ABA receptors are likely to be discovered.

FCA promotes flowering by transcriptionally silencing the floral repressor *FLOWERING LOCUS C* (*FLC*) as well as altering the expression of other genes[5–7]. FCA contains two amino-terminal RNA-binding domains and a carboxy-terminal tryptophan–tryptophan (WW) domain that binds FY, a homologue of a yeast RNA 3′-end processing factor[8]. Razem *et al.*[2] reported that binding of ABA to FCA prevented the interaction with FY, although ABA did not bind directly to the WW domain. Given the reported affinity of ABA for FCA (K_d 19 nM)[2], we expected that low concentrations of ABA would disrupt the FCA–FY interaction. However, FY was still efficiently pulled down by FCA (Fig. 1), even when 5 mM (+)-ABA was added (~1,000 fold molar excess). This shows that ABA does not disrupt the FCA–FY interaction.

To assess directly the ability of FCA to bind ABA, we used [^{3}H]-(±)-ABA. Initially we measured ABA binding to crude protein extracts obtained from 14-day-old *Arabidopsis* seedlings, as reported previously[3]. Although we detected binding to crude extracts (Fig. 2a), no evidence for binding of ABA to two soluble FCA proteins (FCA(ΔN295), residues 296–747, and the more stable FCA(ΔN443), residues 444–747) was obtained (Fig. 2b). We also failed to detect an interaction between unlabelled (+)-ABA and soluble FCA(ΔN443) using isothermal titration calorimetry (data not shown). The conditions used would have detected an interaction with a maximum K_d of 30 μM[9,10]. Re-evaluation of the data reported by Razem *et al.*[2] revealed other inconsistencies. Notably, the FCA protein used by us was produced from the same construct (a gift from the Dean laboratory) that gives rise to a GST–FCA fusion of ~76 kDa (GST is ~27 kDa and FCA(ΔN295) is ~49 kDa); however, Razem *et al.*[2] reported that they used an ~66 kDa protein. Therefore, because 10 μg of protein was used in each assay, the final concentration of protein must have been at least 877 nM, yet binding was saturated after addition of only 50 nM ABA (refer to Fig. 1c in Razem *et al.*[2]). This either indicates that ABA is binding to a minor component or that the concentrations used may have been inaccurately determined.

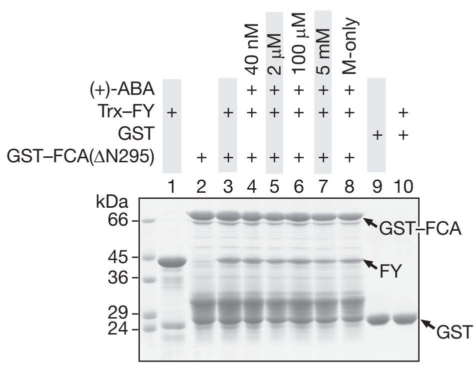


Figure 1 | ABA does not inhibit the FCA–FY interaction. Soluble thioredoxin (Trx)-tagged FY is pulled down by GST–FCA(ΔN295) even in the presence of 5 mM (+)-*cis, trans*-ABA (~1,000-fold molar excess). (+)-ABA was solubilized in methanol. The concentration of ABA used in each sample is indicated; M-only represents addition of methanol alone.

Two other ABA receptors have been reported: the plastid-associated Mg-cheltase H subunit (CHLH)[3] and a protein originally identified as a membrane-bound G-protein-coupled receptor 2 (GCR2)[4]. Because similar methods have been used to characterize the ABA-binding properties of FCA[2], CHLH[3] and GCR2 (ref. 4), our binding data question the role of these proteins as ABA receptors. Our work casts particular doubt over the role of GCR2, because FCA was used as a positive control in experiments showing that GCR2 binds ABA[4]. The possibility that GCR2 may not be an ABA receptor is consistent with recent publications showing that GCR2 is a LanC family member and not a G-protein-coupled receptor[11], and that *gcr2* loss-of-function mutant lines do not have an altered response to ABA[12].

We therefore find no evidence to indicate that FCA is an ABA receptor. An inherent problem with the ABA-binding assay is the inability to remove unbound ABA from samples efficiently. This means that small changes in the total amount bound are difficult to detect accurately and careful data analysis is required. Alternative methods such as isothermal titration calorimetry should be used to validate ABA binding instead.

Note added in proof: The Razem *et al.* paper[2] has been retracted.

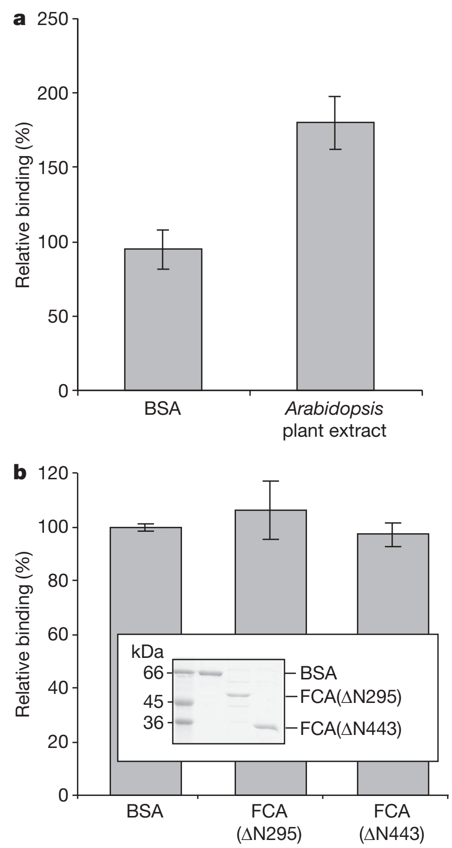


Figure 2 | FCA does not bind ABA. a, ABA binds to *Arabidopsis* plant extract. For assays, 25 μg of crude plant extract[3] and 1 μg of BSA were used. **b**, Recombinant purified FCA does not bind ABA. Assays contained 50 ng of FCA(ΔN443) or FCA(ΔN295); BSA was added so that each sample contained 1 μg of protein. The purified proteins used are shown. For all assays, binding relative to the buffer alone is shown. The background counts (~25% of the total counts added) that remained in buffer-only controls are set at 100%. Error bars indicate mean ± s.d.; $n = 3$, assays were carried out in triplicate using three different protein preparations.

METHODS SUMMARY

Proteins were purified using either a glutathione *S*-transferase (GST) tag (FCA constructs) or a His tag (thioredoxin–FY) after expression in *Escherichia coli*. The GST-pull-down assay was carried out according to Razem *et al.*[2], in sucrose binding buffer[13] (200 µl) at pH 7.0 with 0.5 mM DTT, 0.1% Tween, and (+)-*cis*, *trans*-ABA (AG Scientific, Inc.), as indicated. For washes, sucrose binding buffer contained 0.2% Tween.

ABA-binding assays were as reported[4,14], except the total volume was 50 or 100 µl and unbound [^{3}H]-(±)-ABA (GE Healthcare; 46 Ci mMol^{-1}) was removed with 25 or 50 µl of 0.5% dextran-coated charcoal. To determine binding, three independent protein preparations were assayed in triplicate.

Joanna M. Risk[1], **Richard C. Macknight**[1] **& Catherine L. Day**[1]

[1]Biochemistry Department, University of Otago, Dunedin 9054, New Zealand.

e-mail: catherine.day@otago.ac.nz

Received 25 June; accepted 11 November 2008.

1. Macknight, R. *et al. FCA*, a gene controlling flowering time in *Arabidopsis*, encodes a protein containing RNA binding domains. *Cell* **89**, 737–745 (1997).
2. Razem, F. A., El-Kereamy, A., Abrams, S. R. & Hill, R. D. The RNA-binding protein FCA is an abscisic acid receptor. *Nature* **439**, 290–294 (2006).
3. Shen, Y. *et al.* The Mg-chelatase H subunit is an abscisic acid receptor. *Nature* **443**, 823–826 (2006).
4. Liu, X. *et al.* A G protein-coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid. *Science* **315**, 1712–1716 (2007).
5. Liu, F. *et al.* The *Arabidopsis* RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate *FLC*. *Mol. Cell* **28**, 398–407 (2007).
6. Bäurle, I., Smith, L., Baulcombe, D. C. & Dean, C. Widespread role for the flowering-time regulators FCA and FPA in RNA-mediated chromatin silencing. *Science* **318**, 109–112 (2007).
7. Marquardt, S., Boss, P. K., Hadfield, J. & Dean, C. Additional targets of the *Arabidopsis* autonomous pathway members, FCA and FY. *J. Exp. Bot.* **57**, 3379–3386 (2006).
8. Simpson, G. G., Dijkwel, P. P., Quesada, V., Henderson, I. & Dean, C. FY is an RNA 3′ end-processing factor that interacts with FCA to control the *Arabidopsis* floral transition. *Cell* **113**, 777–787 (2003).
9. Smits, C., Czabotar, P. E., Hinds, M. G. & Day, C. L. Structural plasticity underpins promiscuous binding of the prosurvival protein A1. *Structure* **16**, 818–829 (2008).
10. Majava, V. *et al.* Interaction between the C-terminal region of human myelin basic protein and calmodulin: analysis of complex formation and solution structure. *BMC Struct. Biol.* **8**, doi:10.1186/1472-6807-8-10 (2008).
11. Johnston, C. A. *et al.* Comment on "A G protein coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid". *Science* **318**, 914c, doi:10.1126/science.1143230 (2007).
12. Gao, Y. *et al.* Genetic characterization reveals no role for the reported ABA receptor, GCR2, in ABA control of seed germination and early seedling development in *Arabidopsis*. *Plant J.* **52**, 1001–1013 (2007).
13. Razem, F. A., Luo, M., Liu, J. H., Abrams, S. R. & Hill, R. D. Purification and characterization of a barley aleurone abscisic acid-binding protein. *J. Biol. Chem.* **279**, 9922–9929 (2004).
14. Zhang, D. P., Wu, Z. Y., Li, X. Y. & Zhao, Z. X. Purification and identification of a 42-kilodalton abscisic acid-specific-binding protein from epidermis of broad bean leaves. *Plant Physiol.* **128**, 714–725 (2002).

doi:10.1038/nature07646

Vol 456 | 11 December 2008 | doi:10.1038/nature07525

Brain metabolism dictates the polarity of astrocyte control over arterioles

Grant R. J. Gordon[1], Hyun B. Choi[1], Ravi L. Rungta[1], Graham C. R. Ellis-Davies[2] & Brian A. MacVicar[1]

Calcium signalling in astrocytes couples changes in neural activity to alterations in cerebral blood flow by eliciting vasoconstriction or vasodilation of arterioles. However, the mechanism for how these opposite astrocyte influences provide appropriate changes in vessel tone within an environment that has dynamic metabolic requirements remains unclear. Here we show that the ability of astrocytes to induce vasodilations over vasoconstrictions relies on the metabolic state of the rat brain tissue. When oxygen availability is lowered and astrocyte calcium concentration is elevated, astrocyte glycolysis and lactate release are maximized. External lactate attenuates transporter-mediated uptake from the extracellular space of prostaglandin E_2, leading to accumulation and subsequent vasodilation. In conditions of low oxygen concentration extracellular adenosine also increases, which blocks astrocyte-mediated constriction, facilitating dilation. These data reveal the role of metabolic substrates in regulating brain blood flow and provide a mechanism for differential astrocyte control over cerebrovascular diameter during different states of brain activation.

In the brain, regional changes in neural activity trigger localized alterations in cerebral blood flow[1] through astrocyte activation[2,3]. A rise in intracellular free calcium concentration ($[Ca^{2+}]_i$) within astrocyte endfeet, which collectively circumscribe all cerebral vessels[4], initiates vasoconstriction[5–7] or vasodilation[2,6,8–10]. Cerebral blood flow couples to the lactate/pyruvate ratio and the related nicotinamide adenine dinucleotide (NADH)/NAD^+ ratio[11–13], but links between these metabolic substrates and the mechanisms of cerebral blood flow regulation remain unresolved. We tested the hypothesis that the metabolic state of the tissue, altered by oxygen (O_2) availability, dictates the type of astrocyte influence on arteriole diameter. At the onset of neural activity dendrites rapidly consume O_2 (ref. 14), leading to a reduction in oxygen pressure (p_{O2})[15–17] and oxyhaemoglobin[18,19], which occurs before the increase in cerebral blood flow. Brain metabolism then shifts, whereby glycolysis is enhanced[20,21] in astrocytes[14] and lactate is released[22,23]. Lactate is vasoactive[12,24] and dynamically alters microvasculature diameter in an O_2-dependent manner[25]. Vasodilation occurs in the brain region experiencing the O_2 drop, whereas vasoconstriction ensues in the inactive surrounding area[26]. Delineating the cellular processes responsible for O_2-metabolic effects on cerebral blood flow may be crucial for treating stroke and vascular dementia, as well as aiding in our understanding of neurovascular coupling for the scientific and diagnostic uses of functional magnetic resonance imaging.

p_{O2} converts polarity of vessel response

We tested whether p_{O2} consistently determined if vasoconstrictions or vasodilations were evoked by synaptic activation or by astrocyte Ca^{2+} transients in rat brain slices. In response to periodic 10–20-Hz stimulation of the hippocampal CA3–CA1 pathway in conditions of high O_2 concentration (95%) arterioles constricted (81.4 ± 3.8% (100% = control diameter), $n = 6$, $P < 0.007$), whereas the same vessel in low O_2 concentration (20%) dilated (107.6 ± 0.7%, $n = 6$, $P < 0.005$, $P < 0.0003$ to each other, Fig. 1a–c) in response to the same stimulation. To determine the impact of p_{O2} on astrocyte control of arteriole diameter we applied the metabotropic glutamate receptor (mGluR) agonist (±)-1-aminocyclopentane-trans-1,3-dicarboxylic acid (tACPD) (100 μM), which potently increased astrocyte $[Ca^{2+}]_i$[2,5,8] (240.0 ± 19.1%, $n = 14$, $P < 0.001$, Fig. 2a) and caused arteriole dilation in conditions of low O_2 concentration (107.0 ± 0.8%, $n = 23$, $P < 0.001$, Fig. 2a, f) but constriction in conditions of high O_2 concentration[5] (85.4 ± 4.0%, $n = 10$, $P < 0.006$, Fig. 2f). Next we used two-photon photolysis of the Ca^{2+} cage dimethoxy-nitrophenyl-EGTA-4 (DMNPE-4)[27] to liberate directly astrocyte Ca^{2+} and examine the vessel response. Uncaging DMNPE-4 triggered a Ca^{2+} wave that spread to multiple astrocyte endfeet (Fig. 2b, c). In contrast to previous results in conditions of high O_2 concentration where astrocyte endfeet Ca^{2+} caused constriction[5], in conditions of low O_2 endfeet Ca^{2+} (232.2 ± 8.5%, $n = 17$, $P < 0.001$) caused dilation (107.4 ± 1.0%, $n = 17$, $P < 0.003$, Fig. 2b, c, f). These data indicate that p_{O2} dictates the direction of arteriole diameter change when astrocytes are activated (Supplementary Fig. 2 for p_{O2} measurements).

Vasodilation requires COX and PGE_2

Astrocyte Ca^{2+} activates cytosolic phospholipase A_2, triggering formation of arachidonic acid that is either converted to 20-hydroxyeicosatetraenoic acid (20-HETE) in smooth muscle cells causing vasoconstriction[5] or to the vasodilator prostaglandin E_2 (PGE_2) in astrocytes via cyclooxygenase (COX)[2,9]. We verified COX1 and COX2 expression in astrocytes and their endfeet (Supplementary Fig. 3). The COX inhibitor indomethacin (100 μM) blocked vasodilations caused by tACPD (98.4 ± 0.8%, $n = 9$, $P < 0.0001$ to tACPD alone, Fig. 2f) and by caged Ca^{2+} photolysis (101.0 ± 0.5%, $n = 6$, $P < 0.002$ to uncaging alone, Fig. 2d, f) in conditions of low O_2 concentration. Application of the COX product PGE_2 (1 μM) elicited vasodilation (109.4 ± 2.7%, $n = 4$, $P < 0.05$, Fig. 2e), confirming COX activation and that the generation of PGE_2 is an important signalling molecule in astrocyte-mediated vasodilations in conditions of low O_2 concentration[2,9].

We investigated how changing p_{O2} could modify the type of astrocyte influence on arterioles. Lowering p_{O2} may elevate anaerobic

[1]Brain Research Centre, Department of Psychiatry, University of British Columbia, British Columbia T2N 2B5, Canada. [2]Department of Pharmacology & Physiology, Drexel University College of Medicine, Philadelphia, Pennsylvania 19102, USA.

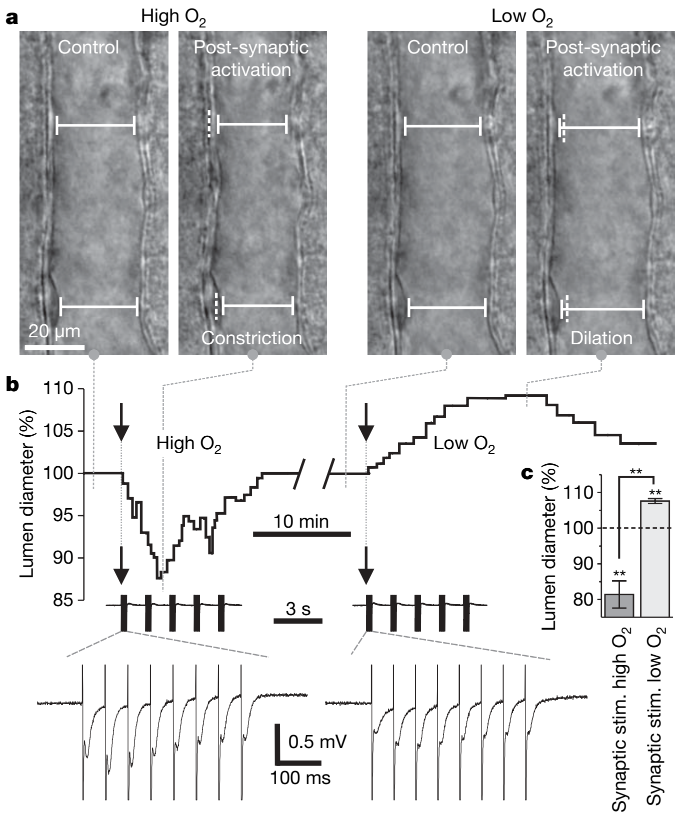


Figure 1 | Lowering p_{O2} converts vasoconstriction to vasodilation. **a**, Arteriole before and after synaptic activation in high O_2 (left) and low O_2 (right). Dashed vertical lines indicate the previous position of the vessel wall. **b**, Top: vessel lumen diameter changes over time in the same vessel shown in **a**. Arrows indicate time of afferent stimulation. Bottom: two expanded timescales show the stimulation protocol (350-ms, 20-Hz train repeated 5 times at 0.75 Hz) and the first train of the field excitatory postsynaptic potentials evoked, verifying synaptic activity. **c**, Summary data ($n = 6$). In all figures, experimental values are the mean ± s.e.m. Double asterisk, $P < 0.01$.

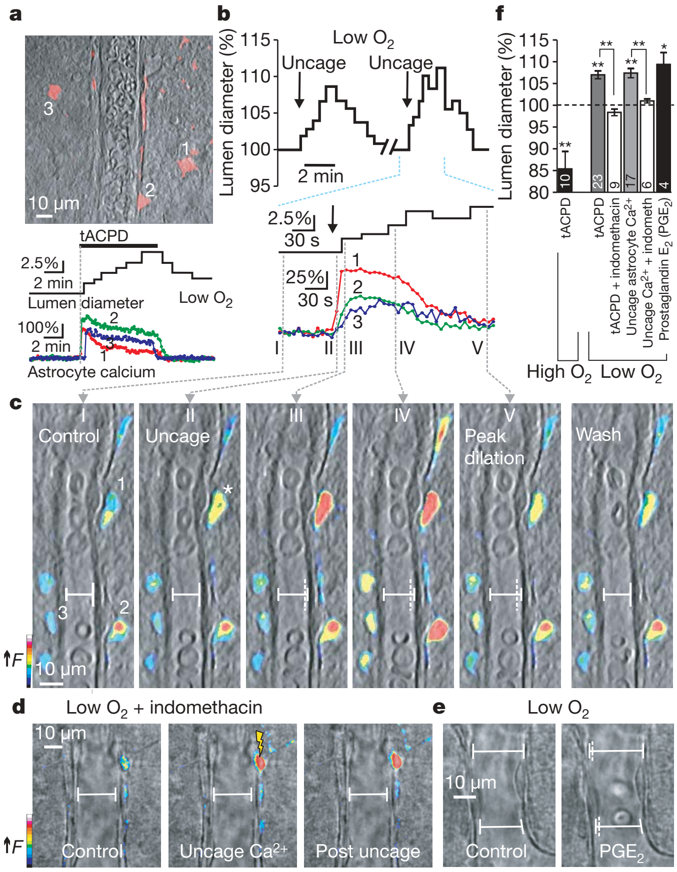


Figure 2 | O_2-mediated conversion to astrocyte-mediated vasodilation requires COX and PGE_2. **a**, Top: astrocytes (red) surround arteriole. Bottom: astrocyte Ca^{2+} signals occur coincident with vasodilation caused by the mGluR agonist tACPD in low O_2. **b**, Top: uncaging astrocyte Ca^{2+} causes vasodilation in low O_2. Bottom: expanded timescale shows that astrocyte endfoot Ca^{2+} precedes diameter increase. **c**, Vessel and pseudo-coloured endfoot Ca^{2+} changes corresponding to times in **b**. **d**, Blocking COX prevents vasodilation caused by uncaging Ca^{2+} in low O_2. **e**, PGE_2 causes vasodilation in low O_2. **f**, Summary data. Asterisk, $P < 0.05$; double asterisk, $P < 0.01$.

metabolism, increasing external lactate. Notably, the action of PGE_2 is terminated when prostaglandin transporters (PGTs) take up PGE_2 by an exchange of intracellular lactate[28]. Owing to the influence of a lactate concentration gradient on PGT efficacy, we tested the hypothesis that higher levels of extracellular lactate reduce PGT uptake of PGE_2, thereby increasing the external PGE_2 concentration resulting in vasodilation. Immunohistochemistry revealed that PGT was expressed in astrocyte endfeet (Fig. 3a) and neurons (Fig. 3b), indicating that both cell types take up PGE_2. Consistent with a positive correlation between external lactate and external PGE_2, low O_2 concentration enhanced both lactate release (low O_2, 114.2 ± 9.1 μM, $n = 6$; high O_2, 41.9 ± 5.6 μM, $n = 6$, $P < 0.001$, Fig. 3c) and extracellular PGE_2 levels when PGE_2 production was triggered by tACPD (low O_2, 136.1 ± 10.2 pg ml^{-1}, $n = 6$; high O_2, 91.9 ± 12.6 pg ml^{-1}, $n = 4$, $P < 0.001$, Fig. 3d). Addition of lactate (1 mM) enhanced the PGE_2 level (control, 40.5 ± 3.3 pg ml^{-1}; with lactate, 58.0 ± 3.1 pg ml^{-1}, $n = 5$, $P < 0.01$) and increased arteriole diameter (107.5 ± 1.0%, $n = 15$, $P < 0.0001$, Fig. 3e–g). Lactate-induced dilations were blocked by indomethacin (100.4 ± 0.4%, $n = 11$, $P < 0.0001$ to lactate alone, Fig. 3g), indicating that they were mediated by PGE_2 via the COX pathway. These data indicate that in low O_2 concentration higher levels of extracellular lactate raise external levels of PGE_2.

Vasodilation requires lactate

We examined the role of astrocyte glycolysis in lactate production[23,29] by imaging the intrinsic fluorescence of the metabolic electron carrier NADH[30]. Two-photon excitation of NADH provides a measure of both oxidative metabolism (punctate mitochondrial fluorescence) and glycolytic metabolism (diffuse cytosolic fluorescence)[14]. We observed that astrocytes, stained with SR-101 (ref. 31), showed diffuse NADH fluorescence in the soma and endfeet (Fig. 3h, i). Stimulating glycolysis with low O_2 increased the astrocyte NADH signal (124.6 ± 1.4%, $n = 5$, $P < 0.003$, Fig. 3i, j) and inhibiting glycolysis with iodoacetate (200 μM) reduced basal NADH levels (87.8 ± 2.0%, $n = 5$, $P < 0.004$, Fig. 3k). Inhibiting lactate dehydrogenase (LDH, which converts pyruvate and NADH to lactate and NAD^+; Supplementary Fig. 1) with oxamate (2.5 mM) increased NADH (119.3 ± 2.0%, $n = 6$, $P < 0.004$, Fig. 3l). These data indicate that astrocyte glycolysis can be augmented by reducing p_{O2}.

Recent two-photon NADH imaging demonstrated an increase in astrocyte glycolysis caused by neuronal activity[14]. We proposed that mGluR activation enhances astrocyte glycolysis in conditions of low O_2, promoting increases in extracellular lactate and vasodilation. tACPD triggered an increase in astrocyte NADH (128.7 ± 4.1%, $n = 7$, $P < 0.0005$, Fig. 4a–e) coincident with lumen widening (108.5 ± 0.7%, $n = 7$, $P < 0.01$, Fig. 4b, c). tACPD enhanced extracellular lactate, which was greatest in low O_2 (low O_2, 186.7 ± 11.2 μM, $n = 6$; high O_2, 98.6 ± 10.2 μM, $n = 7$, $P < 0.001$, Fig. 4f). For additional NADH measurements see Supplementary Figs 4 and 5. We used two pharmacological treatments to limit lactate

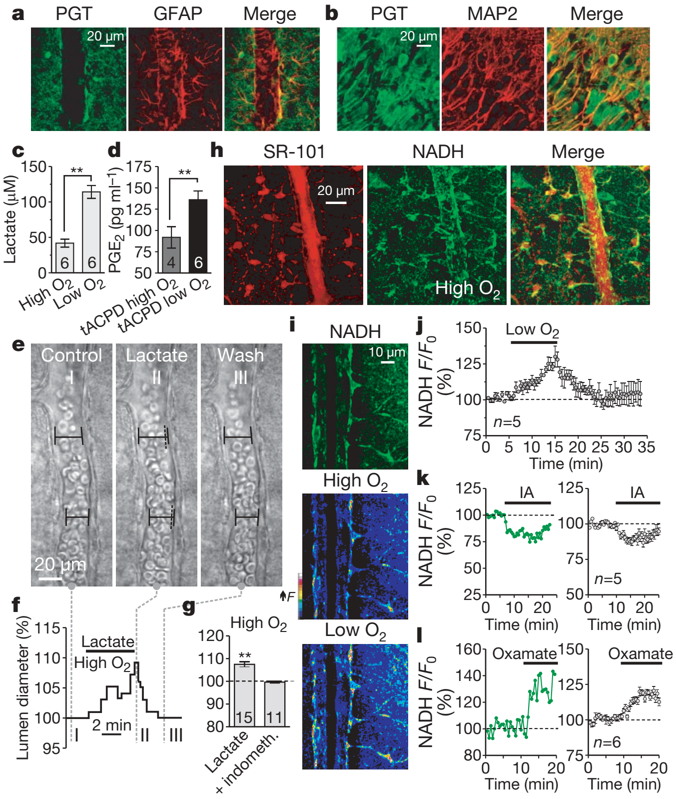


Figure 3 | Low O_2 facilitates lactate and PGE_2 release and enhances astrocyte glycolysis. **a**, **b**, Immunohistochemistry showing that astrocyte endfeet and neurons express PGT. **c**, Lactate release is elevated in conditions of low O_2. **d**, tACPD increases PGE_2 release most in conditions of low O_2. **e–g**, Lactate dilates arterioles and requires COX. Double asterisk, $P < 0.01$. **h**, Metabolic NADH signal (green image stack) in astrocytes (red image stack). **i**, Low O_2 increases NADH signal. **j**, Astrocyte NADH changes caused by low O_2. **k**, **l**, Astrocyte NADH in response to glycolysis inhibition with iodoacetate (IA, **k**) or LDH inhibition with oxamate (**l**). Green traces are single experiments; grey traces are summary data.

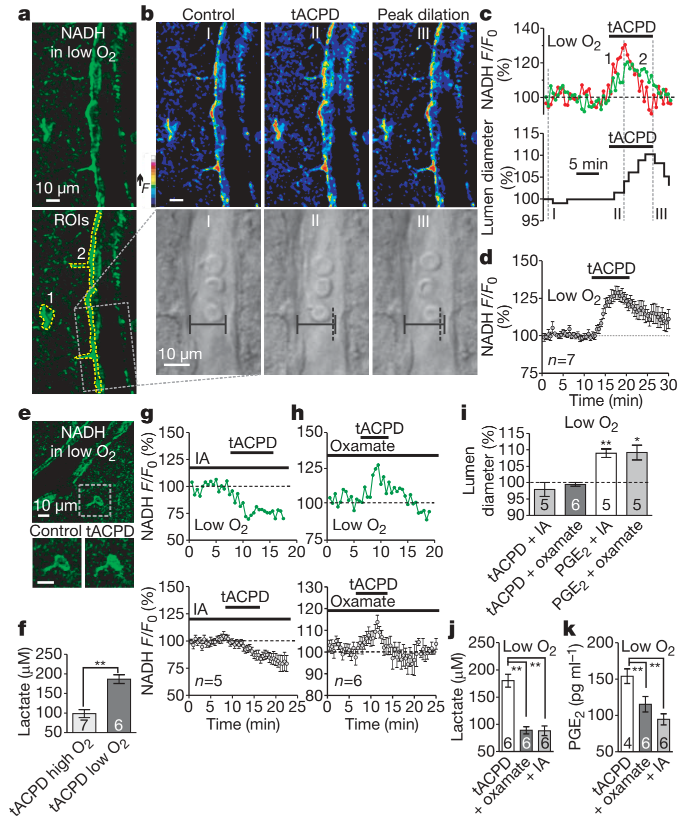


Figure 4 | Glycolysis and lactate release is required for vasodilations. **a–d**, Astrocyte NADH increases (measurements from the regions of interest (ROIs)) coincident with vasodilation in response to tACPD in low O_2. **e**, tACPD increases cytosolic NADH in astrocytes. **f**, tACPD increases extracellular lactate most in conditions of low O_2. **g**, tACPD decreases astrocyte NADH during glycolysis inhibition with iodoacetate (IA). **h**, tACPD increases astrocyte NADH during LDH inhibition with oxamate. Green traces are single experiments; grey traces are summary data. **i**, tACPD fails to dilate vessels in the presence of oxamate or iodoacetate, and PGE_2 rescues vasodilation in these compounds. **j**, **k**, The increase in lactate and PGE_2 levels caused by tACPD is reduced in oxamate and iodoacetate.

release. Using iodoacetate to block the source of lactate, tACPD instead decreased astrocyte NADH ($81 \pm 0.4\%$, $n = 5$, $P < 0.0001$ to tACPD alone, Fig. 4g) and failed to dilate vessels ($97.8 \pm 0.2\%$, $n = 5$, $P < 0.0001$ to tACPD alone, Fig. 4i). The NADH signal decrease was probably due to tACPD-induced astrocyte swelling. Using oxamate to curtail lactate formation, tACPD still increased astrocyte NADH ($112.8 \pm 3.1\%$, $n = 6$, $P < 0.02$, Fig. 4h; see Supplementary Fig. 1), yet vasodilations no longer occurred ($99.0 \pm 0.1\%$, $n = 6$, $P < 0.0001$ to tACPD alone, Fig. 4i). The lack of dilation was associated with reduced extracellular lactate (tACPD, $180.2 \pm 11.9\,\mu M$, $n = 6$; plus iodoacetate, $88.2 \pm 8.8\,\mu M$, $n = 6$, $P < 0.001$; plus oxamate, $89.1 \pm 6.6\,\mu M$, $n = 6$, $P < 0.001$, Fig. 4j) and PGE_2 (tACPD, $154.0 \pm 10.1\,pg\,ml^{-1}$, $n = 4$; plus iodoacetate, $115.4 \pm 10.6\,pg\,ml^{-1}$, $n = 6$, $P < 0.001$; plus oxamate, $94.6 \pm 7.6\,pg\,ml^{-1}$, $n = 6$, $P < 0.01$, Fig. 4k). Because PGE_2 is the final effector molecule on smooth muscle cells, we rescued vasodilation with PGE_2 in iodoacetate ($109.0 \pm 1.3\%$, $P < 0.002$, $n = 5$, Fig. 4i) or in oxamate ($109.2 \pm 2.2\%$, $P < 0.02$, $n = 5$, Fig. 4i and Supplementary Fig. 6). These data demonstrate that inhibition of glycolysis or LDH limits extracellular lactate and PGE_2 accumulation in response to tACPD in low O_2 conditions, preventing astrocyte-mediated vasodilations.

Adenosine blocks vasoconstriction

Our data do not yet explain why vasoconstrictions are absent in low O_2. A low-O_2 environment causes an elevation in extracellular adenosine[32,33], and adenosine A2A receptors reduce smooth muscle cell Ca^{2+} channel activity[34], which normally acts to induce smooth muscle cell contraction. We hypothesized that A2A receptor activation prevents astrocyte-mediated vasoconstriction. Adenosine also suppresses synaptic glutamate release through the A1 receptor[33]; thus, we confirmed an elevated adenosine concentration in low O_2 when the A1 receptor antagonist 8-cyclopentyl-1,3-dipropylxanthine (DPCPX; 100 nM) increased the slope of the extracellular field potential (low O_2, $277.7 \pm 63.6\%$, $n = 9$; high O_2, $99.1 \pm 7.1\%$, $P < 0.05$, $n = 5$, Fig. 5a) and decreased the paired-pulse ratio (low O_2, 1.79 ± 0.04; plus DPCPX, 1.42 ± 0.05, $P < 0.0001$, $n = 9$; high O_2, 1.47 ± 0.01; plus DPCPX, 1.49 ± 0.03, $P > 0.5$, $n = 5$, Fig. 5a) only in low O_2. In the presence of adenosine (100 μM), uncaging astrocyte Ca^{2+} in high O_2 concentration resulted in no vasoconstriction ($99.6 \pm 0.9\%$, $P > 0.6$, $n = 4$, Supplementary Fig. 7). After an initial uncaging event that caused constriction ($87.1 \pm 1.3\%$, $P < 0.002$, Fig. 5b), the A2A receptor agonist CGS21680 (1 μM) blocked subsequent constriction ($99.3 \pm 0.5\%$, $P < 0.0004$ to constriction, $n = 7$, Fig. 5b) despite similar astrocyte Ca^{2+} signals ($F/F_0 = 179.9 \pm 10.9\%$; plus CGS21680 $F/F_0 = 173.8 \pm 10.3\%$, $n = 7$, $P > 0.7$). A2A receptor activation did not affect PGE_2 in response to tACPD ($217.0 \pm 20.4\%$, $n = 5$; plus CGS21680, $186.6 \pm 12.1\%$, $n = 6$, $P > 0.2$), allowing vasodilations to proceed in low O_2. These data demonstrate that increasing extracellular adenosine levels, which occurs in low O_2, prevents astrocyte-mediated vasoconstriction, thereby facilitating the switch to vasodilation.

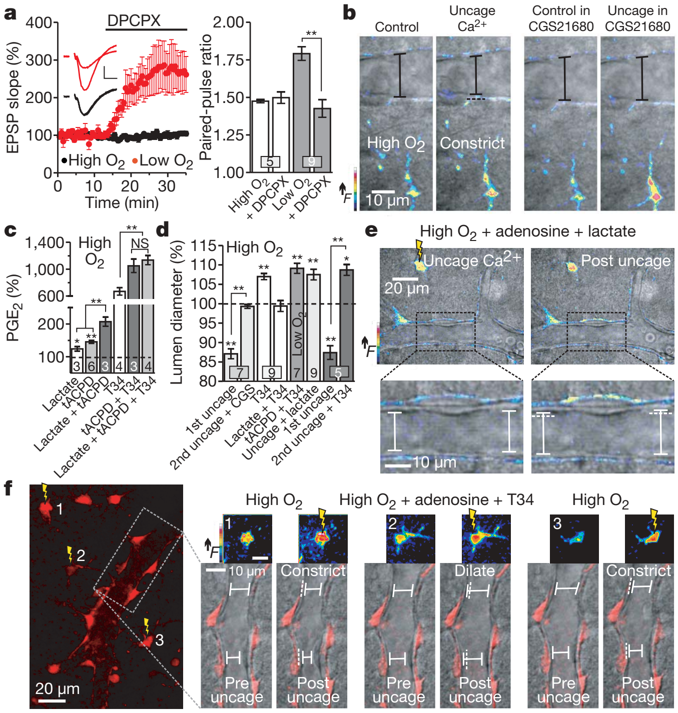


Figure 5 | Raising adenosine and PGE$_2$ levels converts the vessel response. **a**, Enhancement of synaptic transmission by DPCPX shows that extracellular adenosine is elevated in conditions of low O_2 concentration. Scale bars: 0.3 mV, 5 ms. EPSP, excitatory postsynaptic potential. **b**, In conditions of high O_2 concentration uncaging Ca^{2+} fails to cause vasoconstriction in the presence of the A2A receptor agonist CGS21680. **c**, T34 blocks lactate from increasing PGE$_2$ levels. NS, not significant. **d**, Summary of vessel responses. Values for lactate plus T34 and tACPD plus T34 were normalized to T34 baseline. **e**, In conditions of high O_2 concentration uncaging Ca^{2+} in adenosine and lactate triggers vasodilation instead of vasoconstriction. **f**, In conditions of high O_2 concentration vessel responses caused by uncaging Ca^{2+} convert from vasoconstriction to vasodilation in the presence of adenosine and T34.

Lactate attenuates PGT efficacy

We pharmacologically inhibited PGTs to raise extracellular levels of PGE$_2$ and examine the effect on arterioles. Blockade of PGTs with TGBz T34 (ref. 35) (hereafter called T34) (20 μM) elevated PGE$_2$ levels (669.0 ± 110.8%, $P < 0.0001$, $n = 4$, Fig. 5c) and dilated arterioles (107.0 ± 0.8%, $P < 0.001$, $n = 7$, Fig. 5d), indicating tonic PGE$_2$ uptake. Adding tACPD further enhanced the PGE$_2$ level (T34 plus tACPD, 1,048.9 ± 173.6%, $P < 0.02$ compared to T34, $n = 3$, Fig. 5c) and caused further dilation (T34 plus tACPD, 109.1 ± 1.3%, $P < 0.0004$ from T34 baseline, $n = 7$, Fig. 5d), indicating that PGE$_2$ release occurs by diffusion and not by the transporter[36]. However, adding lactate did not further augment PGE$_2$ (T34 plus tACPD plus lactate, 1,134.2 ± 138.7%, $P > 0.4$ compared to T34 plus tACPD, $n = 4$, Fig. 5c) in contrast to the additive effects of lactate and tACPD without T34 (lactate, 124.0 ± 11.9%, $P < 0.02$, $n = 3$; tACPD, 145.6 ± 11.6%, $P < 0.001$, $n = 6$; lactate plus tACPD, 207.0 ± 24.3%, $n = 3$, $P < 0.005$ to tACPD or lactate alone, Fig. 5c). Dilation caused by T34 was not increased further by lactate (T34 plus lactate, 99.4 ± 1.4%, $P > 0.1$ from T34 baseline, $n = 9$, Fig. 5d). These data indicate that T34 occludes the lactate effects of raising PGE$_2$ and causing vasodilation. This demonstrates that PGT efficacy controls the PGE$_2$ level and thus vascular tone.

PGE$_2$ and adenosine convert constrictions

We reasoned that adding lactate (100 μM to 1 mM) and adenosine (100 μM) in conditions of high O_2 would enable astrocyte-mediated constrictions to convert to dilations by attenuating PGE$_2$ uptake and blocking vasoconstriction, respectively. Under this condition, Ca^{2+} uncaging now resulted in dilation (107.5 ± 1.4%, $P < 0.004$ to adenosine alone, $n = 9$, Fig. 5d, e), despite equivalent Ca^{2+} signals in astrocyte endfeet compared with adenosine alone (adenosine, $F/F_0 = 160.4 \pm 6.1\%$, $n = 4$; plus lactate, $F/F_0 = 159.6 \pm 5.1\%$, $P > 0.8$, $n = 9$, Supplementary Fig. 7). We hypothesized that PGT blockade by T34 in high O_2 would have a similar result by reducing PGE$_2$ uptake, thereby raising extracellular PGE$_2$. In the same vessel, an initial uncaging event caused constriction (87.4 ± 1.8%, $P < 0.009$, Fig. 5d, f) but a second uncaging event in the presence of T34 and adenosine caused dilation (108.7 ± 1.4%, $n = 5$, $P < 0.02$, $P < 0.0001$ to constriction, Fig. 5d, f). These data indicate that astrocytes induce vasodilation over vasoconstriction when constrictions are prevented by adenosine and PGE$_2$ uptake is reduced by elevated lactate.

Here we show that by lowering p_{O2}, astrocyte-induced responses convert to vasodilations from vasoconstrictions. Astrocyte Ca^{2+} transients from mGluR activation trigger the synthesis of diffusible arachidonic acid and PGE$_2$. In high O_2, extracellular PGE$_2$ is rapidly cleared by PGTs as a result of low extracellular lactate levels. This keeps extracellular PGE$_2$ levels low and allows astrocyte-derived arachidonic acid to constrict arterioles[5]. In low O_2, astrocyte glycolysis is enhanced and extracellular lactate increases. High external lactate hinders PGE$_2$ clearance and increases extracellular PGE$_2$, which dilates arterioles. In addition, increased adenosine release in low O_2 inhibits astrocyte-mediated vasoconstrictions at the level of smooth muscle cells, blocking the effect of arachidonic acid (Supplementary Fig. 1). Flow through blood vessels is proportional to vessel radius to the fourth power; therefore, a ~9% change in arteriole diameter reported here equates to a ~45% increase in cerebral blood flow. This change is consistent with astrocyte-mediated vasodilations observed *in vivo*[8] and can account for cerebral blood flow changes measured by positron emission tomography[37] and with two-photon microscopy[38,39] during physiological activation. Previous studies on the bidirectional control of vessel diameter implicated different populations of GABA neurons[40], nitric oxide levels[6] and pericytes[41]. Our data indicate that the separate yet competing molecular pathways regulating cerebral blood flow from astrocytes would reach an equilibrium, which shifts depending on the degree of metabolic activity. Under more quiescent periods when O_2 is not being rapidly consumed, astrocyte Ca^{2+} signals induce constrictor tone, keeping cerebral blood flow at an appropriate lower level. During more active periods, the drop in p_{O2} from oxygen consumption and the rise in extracellular lactate and adenosine promote astrocyte-mediated dilation. Manipulating this balance may be a therapeutic avenue for treating the inappropriate declines in cerebral blood flow that occur in some dementias and after stroke.

METHODS SUMMARY

Hippocampal–neocortical slices were prepared from juvenile (postnatal age 16–21 days), male, Sprague–Dawley rats. Treatment of animals was approved by the University of British Columbia Animal Care and Use Committee. Artificial cerebrospinal fluid (ACSF) bubbled with 95% O_2, 5% CO_2, was defined as high O_2 and 20% O_2, 5% CO_2, balanced N_2, was defined as low O_2. Astrocytes were loaded with the Ca^{2+} cage DMNPE-4/AM (10 μM) and/or the Ca^{2+} indicator rhod-2/AM (10 μM) as previously described[5]. A two-photon laser-scanning microscope (Zeiss LSM510-Axioskop-2 fitted with a 40X-W/0.80 numerical aperture objective lens) coupled to a Chameleon ultra-tunable ultra-fast laser (~100-fs pulses 76 MHz, Coherent) provided excitation of rhod-2, intrinsic NADH and was used for uncaging Ca^{2+}. Arterioles were imaged by acquiring the transmitted laser light and using IR-DIC optics. Quantification of lumen diameter, NADH and Ca^{2+} changes were performed with Zeiss LSM (version 3.2) software and ImageJ. The assay used for measuring PGE$_2$ release was Specific Parameter PGE$_2$ ELISA kits (R&D systems). Extracellular lactate levels were measured using a lactate assay kit (Biomedical Research Service Centre). Experimental values are the mean ± s.e.m., expressed as a percentage of 100% control.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 April; accepted 10 October 2008.
Published online 29 October 2008.

1. Mukamel, R. *et al.* Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* **309,** 951–954 (2005).
2. Zonta, M. *et al.* Neuron-to-astrocyte signaling is central to the dynamic control of brain microcirculation. *Nature Neurosci.* **6,** 43–50 (2003).
3. Schummers, J., Yu, H. & Sur, M. Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. *Science* **320,** 1638–1643 (2008).
4. Simard, M., Arcuino, G., Takano, T., Liu, Q. S. & Nedergaard, M. Signaling at the gliovascular interface. *J. Neurosci.* **23,** 9254–9262 (2003).
5. Mulligan, S. J. & MacVicar, B. A. Calcium transients in astrocyte endfeet cause cerebrovascular constrictions. *Nature* **431,** 195–199 (2004).
6. Metea, M. R. & Newman, E. A. Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling. *J. Neurosci.* **26,** 2862–2870 (2006).
7. Chuquet, J., Hollender, L. & Nimchinsky, E. A. High-resolution *in vivo* imaging of the neurovascular unit during spreading depression. *J. Neurosci.* **27,** 4036–4044 (2007).
8. Filosa, J. A., Bonev, A. D. & Nelson, M. T. Calcium dynamics in cortical astrocytes and arterioles during neurovascular coupling. *Circ. Res.* **95,** e73–e81 (2004).
9. Takano, T. *et al.* Astrocyte-mediated control of cerebral blood flow. *Nature Neurosci.* **9,** 260–267 (2006).
10. Filosa, J. A. *et al.* Local potassium signaling couples neuronal activity to vasodilation in the brain. *Nature Neurosci.* **9,** 1397–1403 (2006).
11. Mintun, M. A., Vlassenko, A. G., Rundle, M. M. & Raichle, M. E. Increased lactate/pyruvate ratio augments blood flow in physiologically activated human brain. *Proc. Natl Acad. Sci. USA* **101,** 659–664 (2004).
12. Ido, Y., Chang, K. & Williamson, J. R. NADH augments blood flow in physiologically activated retina and visual cortex. *Proc. Natl Acad. Sci. USA* **101,** 653–658 (2004).
13. Vlassenko, A. G., Rundle, M. M., Raichle, M. E. & Mintun, M. A. Regulation of blood flow in activated human brain by cytosolic NADH/NAD^+ ratio. *Proc. Natl Acad. Sci. USA* **103,** 1964–1969 (2006).
14. Kasischke, K. A., Vishwasrao, H. D., Fisher, P. J., Zipfel, W. R. & Webb, W. W. Neural activity triggers neuronal oxidative metabolism followed by astrocytic glycolysis. *Science* **305,** 99–103 (2004).
15. Vanzetta, I. & Grinvald, A. Increased cortical oxidative metabolism due to sensory stimulation: implications for functional brain imaging. *Science* **286,** 1555–1558 (1999).
16. Ances, B. M., Buerk, D. G., Greenberg, J. H. & Detre, J. A. Temporal dynamics of the partial pressure of brain tissue oxygen during functional forepaw stimulation in rats. *Neurosci. Lett.* **306,** 106–110 (2001).
17. Offenhauser, N., Thomsen, K., Caesar, K. & Lauritzen, M. Activity-induced tissue oxygenation changes in rat cerebellar cortex: interplay of postsynaptic activation and blood flow. *J. Physiol.* **565,** 279–294 (2005).
18. Malonek, D. *et al.* Vascular imprints of neuronal activity: relationships between the dynamics of cortical blood flow, oxygenation, and volume changes following sensory stimulation. *Proc. Natl Acad. Sci. USA* **94,** 14826–14831 (1997).
19. Devor, A. *et al.* Coupling of the cortical hemodynamic response to cortical and thalamic neuronal activity. *Proc. Natl Acad. Sci. USA* **102,** 3822–3827 (2005).
20. Fox, P. T. & Raichle, M. E. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc. Natl Acad. Sci. USA* **83,** 1140–1144 (1986).
21. Fox, P. T., Raichle, M. E., Mintun, M. A. & Dence, C. Nonoxidative glucose consumption during focal physiologic neural activity. *Science* **241,** 462–464 (1988).
22. Hu, Y. & Wilson, G. S. A temporary local energy pool coupled to neuronal activity: fluctuations of extracellular lactate levels in rat brain monitored with rapid-response enzyme-based sensor. *J. Neurochem.* **69,** 1484–1490 (1997).
23. Pellerin, L. & Magistretti, P. J. Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization. *Proc. Natl Acad. Sci. USA* **91,** 10625–10629 (1994).
24. Hein, T. W., Xu, W. & Kuo, L. Dilation of retinal arterioles in response to lactate: role of nitric oxide, guanylyl cyclase, and ATP-sensitive potassium channels. *Invest. Ophthalmol. Vis. Sci.* **47,** 693–699 (2006).
25. Yamanishi, S., Katsumura, K., Kobayashi, T. & Puro, D. G. Extracellular lactate as a dynamic vasoactive signal in the rat retinal microvasculature. *Am. J. Physiol. Heart Circ. Physiol.* **290,** H925–H934 (2006).
26. Devor, A. *et al.* Suppressed neuronal activity and concurrent arteriolar vasoconstriction may explain negative blood oxygenation level-dependent signal. *J. Neurosci.* **27,** 4452–4459 (2007).
27. Ellis-Davies, G. C. Caged compounds: photorelease technology for control of cellular chemistry and physiology. *Nature Methods* **4,** 619–628 (2007).
28. Chan, B. S., Endo, S., Kanai, N. & Schuster, V. L. Identification of lactate as a driving force for prostanoid transport by prostaglandin transporter PGT. *Am. J. Physiol. Renal Physiol.* **282,** F1097–F1102 (2002).
29. Wender, R. *et al.* Astrocytic glycogen influences axon function and survival during glucose deprivation in central white matter. *J. Neurosci.* **20,** 6804–6810 (2000).
30. Chance, B., Cohen, P., Jobsis, F. & Schoener, B. Intracellular oxidation-reduction states *in vivo*. *Science* **137,** 499–508 (1962).
31. Nimmerjahn, A., Kirchhoff, F., Kerr, J. N. & Helmchen, F. Sulforhodamine 101 as a specific marker of astroglia in the neocortex *in vivo*. *Nature Methods* **1,** 31–37 (2004).
32. Frenguelli, B. G., Llaudet, E. & Dale, N. High-resolution real-time recording with microelectrode biosensors reveals novel aspects of adenosine release during hypoxia in rat hippocampal slices. *J. Neurochem.* **86,** 1506–1515 (2003).
33. Brust, T. B., Cayabyab, F. S., Zhou, N. & MacVicar, B. A. p38 mitogen-activated protein kinase contributes to adenosine A1 receptor-mediated synaptic depression in area CA1 of the rat hippocampus. *J. Neurosci.* **26,** 12427–12438 (2006).
34. Murphy, K. *et al.* Adenosine-A2a receptor down-regulates cerebral smooth muscle L-type Ca^{2+} channel activity via protein tyrosine phosphatase, not cAMP-dependent protein kinase. *Mol. Pharmacol.* **64,** 640–649 (2003).
35. Chi, Y., Khersonsky, S. M., Chang, Y. T. & Schuster, V. L. Identification of a new class of prostaglandin transporter inhibitors and characterization of their biological effects on prostaglandin E2 transport. *J. Pharmacol. Exp. Ther.* **316,** 1346–1350 (2006).
36. Chan, B. S., Satriano, J. A., Pucci, M. & Schuster, V. L. Mechanism of prostaglandin E2 transport across the plasma membrane of HeLa cells and *Xenopus* oocytes expressing the prostaglandin transporter 'PGT'. *J. Biol. Chem.* **273,** 6689–6697 (1998).
37. Fox, P. T. & Raichle, M. E. Stimulus rate dependence of regional cerebral blood flow in human striate cortex, demonstrated by positron emission tomography. *J. Neurophysiol.* **51,** 1109–1120 (1984).
38. Kleinfeld, D., Mitra, P. P., Helmchen, F. & Denk, W. Fluctuations and stimulus-induced changes in blood flow observed in individual capillaries in layers 2 through 4 of rat neocortex. *Proc. Natl Acad. Sci. USA* **95,** 15741–15746 (1998).
39. Chaigneau, E. *et al.* The relationship between blood flow and neuronal activity in the rodent olfactory bulb. *J. Neurosci.* **27,** 6452–6460 (2007).
40. Cauli, B. *et al.* Cortical GABA interneurons in neurovascular coupling: relays for subcortical vasoactive pathways. *J. Neurosci.* **24,** 8940–8949 (2004).
41. Peppiatt, C. M., Howarth, C., Mobbs, P. & Attwell, D. Bidirectional control of CNS capillary diameter by pericytes. *Nature* **443,** 700–704 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Murphy, T. Phillips and Y. Tian Wang for reading an earlier version of the manuscript. T34 was a gift from V. L. Schuster and Y. Chi. This work was supported by an operating grant from the Canadian Institutes of Health Research. B.A.M. is a Canada Research Chair and a Michael Smith Foundation for Health Research (MSFHR) Distinguished Scholar. G.R.J.G. is supported by fellowships from the Alberta Heritage Foundation for Medical Research, MSFHR and the Natural Science and Engineering Council of Canada. H.B.C. is supported by postdoctoral fellowships from Wilms Foundation and the Heart and Stroke Foundation of Canada.

Author Contributions G.R.J.G. and B.A.M. designed the imaging experiments and wrote the manuscript. G.R.J.G. performed the imaging experiments and analysis, took slice p_{O2} measurements and tested the effects of synaptic activation on vasomotion. H.B.C., G.R.J.G. and B.A.M. designed the lactate and PGE_2 experiments. H.B.C. and B.A.M. designed the immunohistochemistry experiments. H.B.C. performed the lactate and PGE_2 measurements and analysis and the immunohistochemistry. R.L.R. performed the extracellular field recordings examining adenosine tone. G.C.R.E.-D. designed and synthesized the calcium cage. All authors helped to edit the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.A.M. (bmacvicar@brain.ubc.ca).

METHODS

Slice preparation and O_2 conditions. Sprague–Dawley rats were anaesthetized with halothane (Sigma), decapitated and the brains removed into ice-cold slicing solution containing (in mM): NaCl, 87; KCl, 2.5; $NaHCO_3$, 25; $CaCl_2$, 0.5; $MgCl_2$, 7; NaH_2PO_4, 1.25; glucose, 25; sucrose, 75; saturated with 95% O_2/5% CO_2. Transverse hemi-sections, 400 μm thick, were sliced (Leica vibratome) and incubated at 34 °C in ACSF containing (in mM): NaCl, 126; KCl, 2.5; $NaHCO_3$, 26; $CaCl_2$, 2.0; $MgCl_2$, 1.5; NaH_2PO_4, 1.25; glucose, 10; saturated with 95% O_2/5% CO_2 for 60 min. For experiments slices were at 22–24 °C and perfused at ~2 ml min^{-1}. 20% O_2 provided a p_{O2} on the low end of physiological[16] (Supplementary Fig. 2). Healthy slices can be maintained in 20% O_2 (ref. 42).

Imaging and uncaging. Images were acquired between 50 and 100 μm deep into the slice. Rhod-2 was excited at 835 nm (~3 mW after the objective) and fluorescence was detected with a PMT after passing through a 605-nm (55-nm band-pass) emission filter. For uncaging, the laser was tuned to 730 nm, which also excites rhod-2 (ref. 43). The laser power was carefully increased (~40 mW after the objective) until a Ca^{2+} signal, characteristic of internal release, occurred within astrocytes, triggering a Ca^{2+} wave. The nonlinearity of two-photon microscopy[44] ensures no uncaging occurs during rhod-2 excitation.

For NADH excitation the laser was tuned to 740 nm[14] (~30 mW after the objective) and fluorescence was detected with a PMT after passing through a 450-nm (30-nm band-pass) emission filter. Time series NADH images were acquired every 30 s. NADH *z*-stacks (one control and then one treatment) were acquired continuously. Although exhibiting identical spectral properties to NADH, NADPH is thought to contribute little to the fluorescence signal[45,46].

Lactate and PGE_2 measurements. Protocols in suppliers' instructions were followed for the PGE_2 ELISA and the lactate assays. In both, tetrodotoxin (TTX, 1 μM) and 3,7-dihydro-1-methyl-3-(2-methylpropyl)-1H-purine-2,6-dione (IBMX, 100 μM) were added to dampen neuronal activation and preserve cAMP, respectively. cAMP facilitates glycogen breakdown[47] which is important for astrocyte glycolysis[48]. We confirmed that the same effects occur, yet at higher values, without these compounds.

Immunohistochemistry. Twenty-micrometre sections were used for immunostaining as described[49]. Sections were incubated in blocking solution containing 0.5% BSA and 0.2% Triton X-100 in 0.1 M PBS for 30 min. Sections were incubated overnight at 4 °C with primary antibodies against GFAP (1:1,000; Sigma), MAP2 (1:1,000; Chemicon), COX1 (1:200; Santa Cruz Biotechnology), COX2 (1:200; Santa Cruz Biotechnology), or PGT (1:100; Alpha diagnostic). Sections were then incubated with secondary antibodies: FITC-conjugated mouse anti-rabbit IgG, FITC-conjugated donkey anti-mouse IgG, rhodamine-conjugated rabbit anti-goat IgG or rhodamine-conjugated goat anti-mouse IgG at room temperature for 2 h in darkness (all secondary antibodies from Santa Cruz Biotechnology).

Electrophysiology. For extracellular recordings examining adenosine, slices were maintained at 34 °C and bubbled with an O_2 treatment 45 min before experimentation. Field excitatory postsynaptic potentials (fEPSPs) were evoked by stimulating the hippocampal CA3–CA1 pathway using a concentric bipolar electrode (Frederick Haer Co.) and were acquired in stratum radiatum. fEPSPs were evoked every 30 s and were analysed for slope (mV ms^{-1}) (see ref. 33 for details). Responses were normalized to the control mean. A second response, 50 ms after the first, was evoked for the paired-pulse. The paired-pulse ratio equals the average of ten sweeps during DPCPX/average of ten sweeps immediately before DPCPX.

To test synaptic activation in vasomotion, a stimulating electrode and recording electrode were positioned ~250 μm apart in the CA3–CA1 pathway, straddling the vessel. A 10–20-Hz stimulation train 350–1,000 ms in duration was repeated 5–10 times at a rate of 0.5–1 Hz. Once a constriction was observed in high O_2, the same stimulation protocol was used in low O_2 to evoke dilation. Forty minutes equilibration was allowed when the p_{O2} was changed.

Data collection, analysis and statistics. An image (512 × 512 pixels) was collected in 393.2–983.4 ms, using 8-line averaging. Lumen diameter measurements were made at multiple sites along the vessel. Fluorescence signals were defined as F/F_0 (%) $= [(F_1 - B_1)/(F_0 - B_0)]100$, where F_1 and F_0 are fluorescence at a given time and the control period mean, respectively. B_1 and B_0 are the corresponding background fluorescence signals. Background values were taken from the neuropil for rhod-2 and the vessel lumen for NADH. Pseudo-colour images show absolute changes in fluorescence (ImageJ, 16-colour linear Lut). Experimental values are the mean ± s.e.m.; baseline equals 100%; *n* is the number of experiments conducted. Lumen diameter changes: 90% means a reduction in diameter by 10% and 110% means an increase in diameter by 10%. Statistical tests were either a two-tailed Student's *t*-test or an ANOVA with a Neumann–Keuls post-hoc test for comparison between multiple groups. $P < 0.05$ was accepted as statistically significant (asterisk, $P < 0.05$; double asterisk, $P < 0.01$).

Drugs. Sigma supplied tACPD, applied for 5–10 min; lactate, applied for 5–10 min and pH-corrected to 7.4; sodium-oxamate and sodium-iodoacetate, applied continuously with 15 min incubation (iodoacetate exposure of ~30 min generated an 'anoxic-depolarization-like' wave causing vessel spasms; tACPD was assessed before this event); indomethacin, applied continuously with 20 min incubation; IBMX, present during the release assays; U46619, applied continuously with 20 min incubation. Tocris supplied DPCPX, applied continuously; CGS21680, applied continuously with 20 min incubation. Alamone labs supplied tetrodotoxin, present during release assays. Cayman Chemicals supplied PGE_2, applied for 5 min. Invitrogen supplied rhod-2/AM. DMNPE-4/AM was synthesized by G. Ellis-Davies. T34, puffed from a pipette or bath applied, was a gift from V. L. Schuster.

42. D'Agostino, D. P., Putnam, R. W. & Dean, J. B. Superoxide ($\bullet O_2^-$) production in CA1 neurons of rat hippocampal slices exposed to graded levels of oxygen. *J. Neurophysiol.* **98,** 1030–1041 (2007).
43. Xu, C., Zipfel, W., Shear, J. B., Williams, R. M. & Webb, W. W. Multiphoton fluorescence excitation: new spectral windows for biological nonlinear microscopy. *Proc. Natl Acad. Sci. USA* **93,** 10763–10768 (1996).
44. Denk, W. Two-photon scanning photochemical microscopy: mapping ligand-gated ion channel distributions. *Proc. Natl Acad. Sci. USA* **91,** 6629–6633 (1994).
45. Klaidman, L. K., Leung, A. C. & Adams, J. D. Jr. High-performance liquid chromatography analysis of oxidized and reduced pyridine dinucleotides in specific brain regions. *Anal. Biochem.* **228,** 312–317 (1995).
46. Vishwasrao, H. D., Heikal, A. A., Kasischke, K. A. & Webb, W. W. Conformational dependence of intracellular NADH on metabolic state revealed by associated fluorescence anisotropy. *J. Biol. Chem.* **280,** 25119–25126 (2005).
47. Sorg, O. & Magistretti, P. J. Characterization of the glycogenolysis elicited by vasoactive intestinal peptide, noradrenaline and adenosine in primary cultures of mouse cerebral cortical astrocytes. *Brain Res.* **563,** 227–233 (1991).
48. Brown, A. M. & Ransom, B. R. Astrocyte glycogen and brain energy metabolism. *Glia* **55,** 1263–1271 (2007).
49. Ryu, J. K. *et al.* Microglial activation and cell death induced by the mitochondrial toxin 3-nitropropionic acid: *in vitro* and *in vivo* studies. *Neurobiol. Dis.* **12,** 121–132 (2003).

Vol 456 | 11 December 2008 | doi:10.1038/nature07585

ARTICLES

Antigenic variation in *Giardia lamblia* is regulated by RNA interference

César G. Prucca[1]*, Ileana Slavin[1]*, Rodrigo Quiroga[1], Eliana V. Elías[1], Fernando D. Rivero[1], Alicia Saura[1], Pedro G. Carranza[1] & Hugo D. Luján[1]

Giardia lamblia (also called _Giardia intestinalis_) is one of the most common intestinal parasites of humans. To evade the host's immune response, _Giardia_ undergoes antigenic variation—a process that allows the parasite to develop chronic and recurrent infections. From a repertoire of ~190 variant-specific surface protein (VSP)-coding genes, _Giardia_ expresses only one VSP on the surface of each parasite at a particular time, but spontaneously switches to a different VSP by unknown mechanisms. Here we show that regulation of VSP expression involves a system comprising RNA-dependent RNA polymerase, Dicer and Argonaute, known components of the RNA interference machinery. Clones expressing a single surface antigen efficiently transcribe several VSP genes but only accumulate transcripts encoding the VSP to be expressed. Detection of antisense RNAs corresponding to the silenced VSP genes and small RNAs from the silenced but not for the expressed _vsp_ implicate the RNA interference pathway in antigenic variation. Remarkably, silencing of Dicer and RNA-dependent RNA polymerase leads to a change from single to multiple VSP expression in individual parasites.

Clonal phenotypic variation of surface-exposed antigenic determinants (antigenic variation) is a major evasion process developed by several microorganisms, which use different mechanisms to switch the expression of their surface antigens to maintain chronic infections under the continuous immune pressure generated by their hosts[1]. The protozoan *G. lamblia* also undergoes antigenic variation[2,3]. This binucleated, flagellated parasite not only is one of the most common causes of human intestinal disease but also is of great biological interest because it derives from the earliest branch of the eukaryotic line of descent[2,4]. In *Giardia*, antigenic variation accounts for the variable and/or persistent course of some infections as well as for the propensity for multiple reinfections, and involves a family of VSPs[2,3]. VSPs cover the entire surface of the trophozoite and are the main antigens recognized by the host immune system[3]. VSPs vary in size from 30 kDa to 200 kDa; they possess a variable cysteine-rich amino-terminal region and a conserved carboxy-terminal region that includes a hydrophobic transmembrane domain and a short cytosolic tail[3]. The parasite genome encodes a repertoire of about 190 VSP genes[4], but only one VSP is expressed at any given time on the surface of individual trophozoites[2–8]. Switching to the expression of another VSP occurs once every 6–13 generations, even in the absence of any immunological pressure[7]. Similar to most other *G. lamblia* genes, VSP genes have no introns and their 5′ upstream regions are relatively short and have been found to have limited or no sequence conservation. Moreover, there are no typical eukaryotic promoters present in these regions. The 3′ downstream regions from *Giardia* genes, including VSP-coding genes, also tend to be short, typically 0–30 nucleotides long. So far, neither gene-rearrangement processes nor promoter-dependent switch-on/switch-off mechanisms have been demonstrated to be involved in *Giardia*'s antigenic switching[2,3,7].

To delve into the molecular mechanisms involved in *Giardia*'s antigenic variation, we initially performed nuclear run-on experiments to determine whether regulation of VSP expression is controlled at the transcriptional or at the post-transcriptional level. Then, we searched for the existence of sense and antisense VSP RNAs in trophozoites by polymerase chain reaction with reverse transcription (RT–PCR) and for the activity of enzymes involved in the synthesis and degradation of double-stranded (ds)RNAs in some eukaryotes, such as RNA-dependent RNA polymerase (RdRP), Dicer and Argonaute. The characterization involved cloning and expressing these genes as well as the analysis of small RNAs generated from VSP dsRNAs. Additionally, expression of different VSPs was evaluated after silencing the components of the *Giardia* RNA interference (RNAi) pathway.

Transcription of VSPs in *Giardia*

Transcription of VSP genes was analysed by nuclear run-on assays using RNA isolated from trophozoites of clone WB9B10 (Fig. 1a), which only express VSP9B10 (GenBank accession number AAK97086) on their surface (Supplementary Fig. 1). The results indicate that most of the VSP genes were simultaneously transcribed. In contrast, when total RNAs extracted from two different *Giardia* clones (WB9B10 and WB1267) were probed with an oligonucleotide corresponding to the conserved 3′ end of VSP genes, only one transcript of the molecular size corresponding to the VSP expressed on the surface of these clones was detected (Fig. 1b). Additionally, very low-molecular-weight bands, suspected to be degradation products, were seen. Accumulation of only one VSP transcript was observed in different *Giardia* clones using *vsp*-specific probes (reviewed in ref. 3). These results raise the question of why only one VSP transcript accumulates if more than one *vsp* is transcribed at the nuclei of the parasite. One possible explanation is that the non-traduced VSP messenger RNAs are specifically degraded, a feature of a post-transcriptional gene silencing (PTGS) model[9,10].

Post-transcriptional gene silencing in *Giardia*

A key step of PTGS is the production of dsRNAs that share homology with the silenced gene[9,10]. RT–PCR assays designed to specifically amplify sense or antisense VSP products revealed, after cloning and sequencing these fragments, that RNAs of both strands are present in trophozoites (Supplementary Fig. 2). To evaluate the possible simultaneous transcription of sense and antisense RNAs for VSP-coding

[1]Laboratorio de Bioquímica y Biología Molecular, Facultad de Medicina, Universidad Católica de Córdoba, Córdoba X5004ASK, Argentina.
*These authors contributed equally to this work.

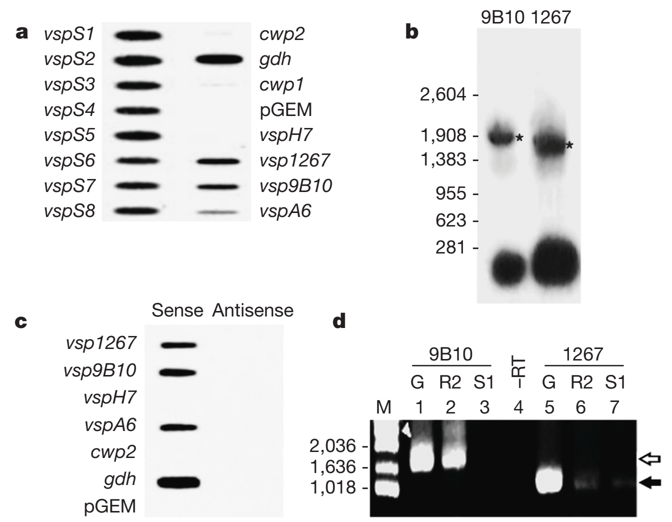


Figure 1 | Several VSP genes are simultaneously transcribed in *Giardia*. **a**, Nuclear run-on using freshly isolated *Giardia* nuclei transcribed *in vitro* in the presence of [^{32}P]UTP. Total RNA from clone WB9B10 was used to probe a membrane on which different DNA products were slot blotted. Isolated *vsp* fragments (*vspS1–S8*, Supplementary Fig. 2) are on the left, whereas specific genes (*vsp9B10*, *vsp1267*, *vspA6*, *vspH7*, *cwp1*, *cwp2* and *gdh*) and the cloning vector pGEM, used as negative controls, are on the right. *cwp1* and *cwp2* were not transcribed in non-encysting trophozoites, and neither was *vspH7*, which is not present in the genome of the WB isolate, whereas *gdh* showed active transcription. **b**, Northern blot analysis on total RNA extracted from clones WB9B10 and WB1267 probed with the 3′ conserved region of *vsp* (Supplementary Fig. 2). In addition to very-low-molecular-weight species, a single transcript of the size corresponding to the mRNA encoding the clone-specific VSP was found (asterisks). Size markers in nucleotides are displayed on the left. **c**, Sense and antisense *in-vitro*-transcribed genes (*vsp9B10*, *vsp1267*, *vspA6*, *vspH7*, *cwp2* and *gdh*) were slot blotted and hybridised with the products of nuclear run-on under the same conditions as in Fig. 1a. Only sense transcripts were detected. **d**, Comparison among PCR products generated from clone WB9B10 trophozoites by specific primers (9B10F/9B10R and 1267F/1267R) on genomic DNA (lanes 1 and 5), or cDNAs generated with either reverse primer R2 (lanes 2 and 6) or sense primer S1 (lanes 3 and 7). Lane 4 is a control of 7 without RT. The white arrow indicates the *vsp9B10* fragment, which is present in genomic DNA and in the sense cDNA but not in the antisense cDNA. The black arrow indicates the *vsp1267* fragment, which is present in genomic DNA, in the sense cDNA and in the antisense cDNA. M, molecular marker.

genes, as suggested to occur for other genes in *Giardia* by a relaxed mechanism controlling transcription[11], we performed a second nuclear run-on experiment using specific sense and antisense probes (Fig. 1c). In this assay, we were unable to detect VSP-antisense RNAs, indicating that those molecules could be generated post-transcriptionally. We also analysed PCR products generated from clone WB9B10 using *vsp9B10*- and *vsp1267*-specific primers (Fig. 1d). The band corresponding to *vsp9B10* was present in genomic DNA and sense complementary DNA but poorly in antisense cDNA. In contrast, *vsp1267* (GenBank accession number M63966), which is not expressed on the surface of clone WB9B10, could be amplified from genomic DNA, and equally amplified from both sense and antisense cDNAs. These results demonstrate that VSPs are transcribed simultaneously (further supporting the results of the nuclear run-on experiments), and that there is an low abundance of antisense transcripts for the VSP that is expressed and a presence of antisense RNAs for VSPs that are transcribed but not translated.

Components of the *Giardia* RNAi machinery

RdRP-mediated, unprimed production of dsRNAs from aberrant mRNAs and primed/unprimed production of dsRNA guided by short interfering RNAs (siRNAs) is necessary for triggering RNAi in some organisms[12–14]. It has been suggested that this enzyme could be acting as 'an aberrant-RNA sensor'[15]. A *Giardia* homologue of RdRP was identified. This RdRP gene encodes a basic protein of 155,257 Da that shares high homology with other eukaryotic RdRPs and greatly differs from the one encoded by the *Giardia* virus[16], indicating the protozoal nature of the identified gene. Analysis of the *Giardia* RdRP sequence indicates that this protein lacks signal peptide and transmembrane domains, but contains a ribosomal protein S2 signature-1 motif, which has been associated with mRNA binding[17]. *Giardia* RdRP transcription was verified by RT–PCR and northern blotting, and its localization assessed by haemagglutinin-tagged expression. RdRP of different organisms have cytosolic or nuclear localization, but our results show that *Giardia* RdRP is probably associated to ribosomes present on the cytoplasmic side of the endoplasmic reticulum. Moreover, the enzyme is active in trophozoites, because it was capable of forming high-molecular-weight RNAs *in vitro* in the presence of homologous VSP RNAs (Supplementary Fig. 3b).

A characteristic of RNAi is the degradation of dsRNA into ~21–25-nucleotide siRNAs by the dsRNA-specific RNase Dicer[18]. Previously, a *Giardia* Dicer homologue was identified, its structure solved, and the *in vitro* Dicer activity of the recombinant protein demonstrated[19]. Here we show that this enzyme is constitutively expressed during the entire life cycle of *Giardia* and verify its cytoplasmic localization (Supplementary Fig. 3). To evaluate the activity of *Giardia* Dicer, *in vitro* assays using radiolabelled dsRNA were exposed to a post-nuclear *Giardia* extract. Our results (Fig. 2a) demonstrated that, regardless of the gene and the strand that is labelled (sense, antisense, or both), dsRNAs are processed into small RNA fragments of 20–30 nucleotides (similar to those shown for the recombinant enzyme in ref. 19), and that this processing is favoured, as in other eukaryotes[20], by the presence of ATP (Fig. 2b). In addition, small RNAs obtained from those experiments were able to be cloned similar to siRNAs that have 5′-P and 3′-OH ends. Sequencing of those siRNAs indicated that they derived from the input VSP genes and that they are 22–25 nucleotides in length (Supplementary Table 1).

Argonaute proteins are essential players in gene-silencing pathways guided by small RNAs. These proteins are part of the RNA-induced silencing complex that recognizes the target mRNA and guide post-transcriptional gene silencing either by destabilization of the mRNA or by translational repression[21]. We were able to identify in *Giardia* a single Argonaute protein (*Giardia* AGO) that has PIWI and PAZ domains. Its expression was evaluated by northern blotting and its cellular localization determined by expressing an haemagglutinin-tagged version of the protein. Similar to that observed in other cells[21], *Giardia* AGO localizes to the cytoplasm (Supplementary Fig. 3).

Regulation of VSP expression

The occurrence of multiple, homologous VSP transcripts could direct the generation of antisense RNAs by *Giardia* RdRP after transcription of several VSP genes took place. Moreover, the presence and activity of Dicer, and probably of AGO, suggests that an RNAi-like mechanism might be involved in regulation of the expression of surface antigenic variants in *Giardia*. In this cell in particular, a post-transcriptional regulatory system lying in the cytoplasm might be beneficial to synchronize both transcriptionally active nuclei of the parasite[2].

To test this hypothesis, we analysed whether the *Giardia* PTGS machinery could discriminate the presence of different VSP mRNAs by mixing *Giardia* cytoplasmic extracts with one, two or three different VSP transcripts generated *in vitro*. When two or more labelled VSP mRNAs were incubated with trophozoite extracts containing the RNAi machinery, small VSP RNAs were produced, with an identical pattern to the Dicer activity assay products (compare Fig. 2a and Fig. 2c). In contrast, whenever a single transcript was incubated, no mRNA degradation took place. Additionally, if the unrelated genes *cwp2* (which encodes the cyst wall protein 2; not shown) or *gdh* (which encodes glutamate dehydrogenase) were added to a single radiolabelled VSP mRNA, no degradation to small

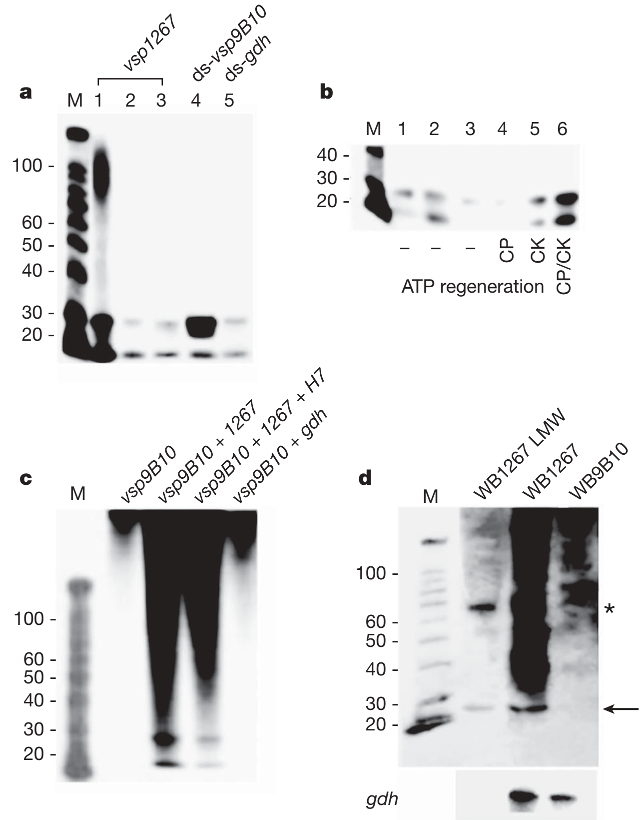


Figure 2 | Dicer activity and detection of VSP small RNAs in *Giardia*. **a**, Generation of small RNAs from dsRNAs by *Giardia* extracts demonstrated Dicer-like activity. For *vsp1267* dsRNA, both strands (lane 1), only the sense RNA strand (lane 2) or only the antisense RNA strand (lane 3) were radioactively labelled. For *vsp9B10* and *gdh* both strands were labelled. Double-stranded RNAs were incubated with a *Giardia* WB9B10 extract at 37 °C for 1 h; next, total RNA was isolated and electrophoresed. In all cases, small RNAs were produced. **b**, Effect of the presence of ATP on dsRNA processing by *Giardia*. Lanes 1 and 2 show untreated controls: incubation of *vsp1267* dsRNA with *Giardia* cytoplasmic lysates without ATP depletion (1 h and 3 h of incubation, respectively). ATP was depleted with 2 mM glucose and 0.1 U μl^{-1} hexokinase (lane 3) and then regenerated using phosphocreatine (CP, lane 4), creatine kinase (CK, lane 5) or both (lane 6). **c**, Generation of small RNAs by incubation of *vsp* riboprobes with *Giardia* extracts from clone WB9B10. One, two or three different VSP mRNAs (*vsp9B10*, *vsp1267*, *vspH7*) were mixed and incubated with *Giardia* extracts. [^{32}P]-labelled small RNAs were produced whenever more than one transcript was present. RNA size markers in nucleotides are on the left. **d**, Total RNA from clone WB9B10 and clone WB1267 trophozoites as well as low-molecular-weight (LMW) RNAs from the clone WB1267 were electrophoresed, blotted and probed using partially digested *in-vitro*-transcribed *vsp9B10* RNA. In clone WB9B10, no small RNAs were found. In contrast, small RNAs from *vsp9B10* were present in clone WB1267 (arrow), which does not express VSP9B10. Interestingly, ~70-nucleotide-long RNAs (asterisk) that might represent incompletely digested mRNAs can also be observed. *gdh* was used as a loading control.

RNAs was detected (Fig. 2c), indicating that the silencing machinery specifically processes homologue RNAs. Similar results were obtained when using cell extracts from different *Giardia* clones (WBA6, WB1267 or GSH7): *vsp9B10* RNA was processed to small RNAs only when combined with other homologue VSP genes but not when it was the only *vsp* added to the reaction (Supplementary Fig. 4). It is clear that, under the conditions used in these experiments (excess of exogenous RNAs, short incubation times), the presence in the trophozoite extracts of endogenous siRNAs and various sense and antisense VSP RNAs certainly does not interfere with the silencing process. Because the VSP mRNAs used in these experiments were synthesized *in vitro*, it is obvious that the silencing mechanism can discriminate among homologous mRNAs in the absence of any possible post-transcriptional RNA modification.

Considering that sense and antisense VSP transcripts were found in *Giardia* and that Dicer activity was demonstrated experimentally, we looked for small RNAs resulting from *vsp* dsRNA degradation. Using partially a digested *vsp9B10* probe for northern blot assays in *Giardia* clones WB9B10 and WB1267, we were able to detect small RNAs for VSP9B10 in the clone that does not express that VSP (WB1267 in this case). On the other hand, no small RNAs for VSP9B10 were found in clone WB9B10, which expresses VSP9B10 on the parasite's surface (Fig. 2d)

These results raise the question of how a single VSP transcript bypasses this silencing process and is translated and expressed at the surface of the parasite. In some organisms it is known that RNAi correlates with methylation of homologous DNA sequences, which in turn modulates transcription[22]. Nonetheless, DNA methylation has been reported to be absent in this parasite[2]. Moreover, methylated nucleotides could not be detected in *Giardia* DNA by high-performance liquid chromatography (Supplementary Fig. 5). In this context, it is reasonable that after the transcription of the *vsp* repertoire, differences in concentration of each VSP transcript might constitute an exclusion factor that allows the temporally most abundant VSP mRNA to saturate the silencing machinery and be expressed at the parasite surface.

To test whether the concentration of different VSP transcripts has a role in antigenic switching, we unbalanced VSP expression *in vivo* by expressing in clone WB9B10 either *vspH7* (GenBank accession number, AAA18202) or constructs containing antisense regions of *vsp9B10* under the control of a strong promoter such as that of the α-tubulin gene[23,24]. VSPH7 showed variable expression even under the pressure of the selection drug, and VSP9B10 also decreased over the time (Supplementary Fig. 6). Besides, when VSP9B10 was knocked down, its expression on the parasite's surface diminished slightly faster than in the control (Supplementary Fig. 7). These results suggest that the promoter region of *vsp*s has little or no influence on VSP expression and, therefore, that a PTGS mechanism must be involved in *Giardia*'s antigenic variation. Additional *in vitro* experiments in which the concentration of different VSP transcripts was unbalanced demonstrated that mRNA concentration may be relevant for a given VSP to circumvent the silencing machinery and be translated (Supplementary Fig. 8).

Silencing of *Giardia* RdRP, Dicer and AGO

Because specific gene knockouts are not possible in *Giardia* owing to its polyploidy and the presence of two nuclei[2], we sought for direct evidence supporting the involvement of the characterized RNAi components during antigenic variation by knocking down the expression of *Giardia dicer, rdrp* and *ago* by constitutive expression in trophozoites of part of their antisense transcripts[23]. When a reduction of the expression of *rdrp* (RdRP-AS) or *dicer* (Dicer-AS) took place (Supplementary Fig. 9), trophozoites that express more than one VSP in their surface were generated, as determined by immunofluorescence assays using specific monoclonal antibodies (Fig. 3a and Table 1), flow cytometry (Fig. 3b) and western blotting (Fig. 3c). Silencing of *Giardia ago* did not produce any viable clone, suggesting that this molecule is essential for the parasite, probably by maintaining the integrity of the genome by controlling the action of transposons[25]. Cells in which Dicer or RdRP were knocked down proliferate and encyst in culture as normal, and no deleterious effect on VSP regulation caused by the procedure was observed (Table 1).

Our results are consistent with a major role for the cellular PTGS machinery itself in the selection of one VSP transcript destined to be expressed in a single trophozoite. Although this PTGS mechanism resembles gene silencing phenomena controlling the expression of foreign genes[12], in *Giardia* it targets a family of endogenous genes, as recently demonstrated in mouse oocytes[26,27]. It has been shown that antisense transcripts for CWP1 also appear early during *Giardia*

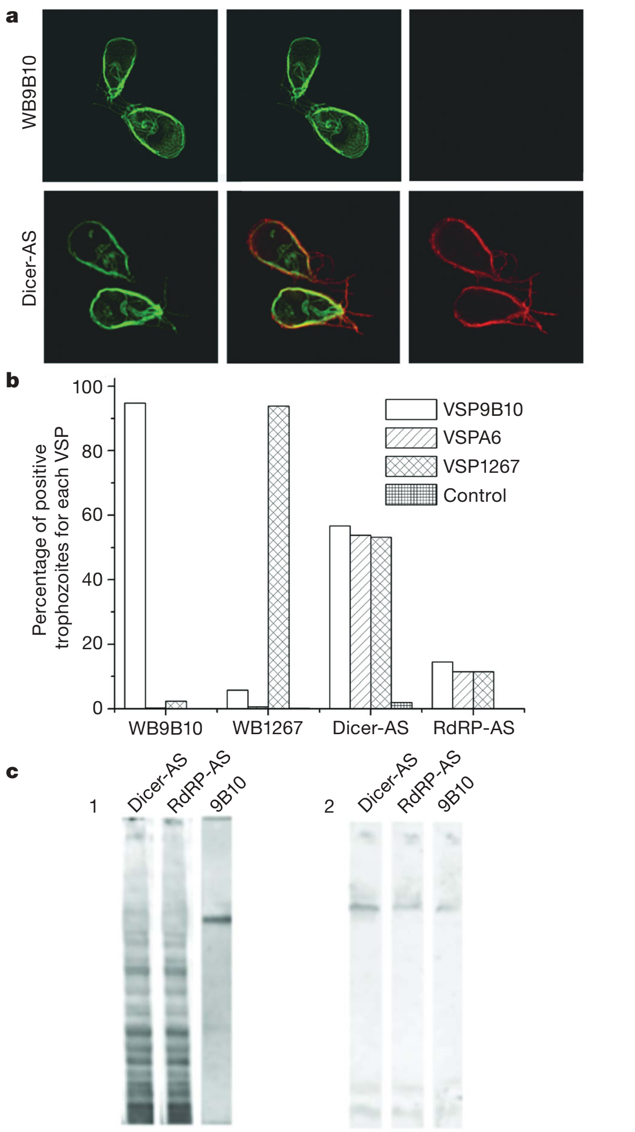


Figure 3 | Expression of different VSPs in *Giardia rdrp*- and *dicer*-knockdown transgenic trophozoites. **a**, Direct immunofluorescence assay on mock-transfected (top panel) or Dicer-AS-transfected (bottom panel) trophozoites with FITC-conjugated monoclonal antibody 9B10 (VSP9B10; left panel) and TRITC-conjugated monoclonal antibody 5C1 (VSP1267; right panel). When *Giardia* Dicer expression was knocked down, trophozoites expressing VSP9B10 on the surface also expressed VSP1267 (merged image; middle panel). **b**, Percentage of *Giardia* trophozoites expressing a particular VSP as determined by flow cytometry assays using specific monoclonal antibodies (VSP9B10, monoclonal antibody 9B10; VSP1267, monoclonal antibody 5C1; VSPA6, monoclonal antibody 6A7) on clone WB9B10, clone WB1267, and with cells transfected with antisense constructs for *Giardia* RdRP (RdRP-AS) or *Giardia* Dicer (Dicer-AS). Goat anti-mouse immunoglobulins were used as a negative control. Clones display clear single-VSP expression patterns, whereas RdRP-AS and Dicer-AS show expression of more than one VSP per trophozoite. **c**, Western blot analysis of protein extracts from wild-type WB9B10 and WB9B10 trophozoites in which Dicer has been knocked down. After electrophoresis and transfer to nitrocellulose, the filters were hybridised with either (1) the monoclonal antibody 12F1 clone G3 (generated against the CRGKA domain conserved in all VSPs) or (2) monoclonal antibody 9B10 (specific for VSP9B10). Results show that the specific monoclonal antibody anti-VSP9B10 recognizes only one band in clone WB9B10 and in transgenic trophozoites, whereas the monoclonal antibody 12F1 reacts with many protein species in the transgenic cells, indicating that many VSPs are expressed simultaneously in *Giardia* when the RNAi pathway has been disrupted.

Table 1 | Quantitative analysis of VSP expression in *Giardia rdrp*- and *dicer*-knockdowns

VSP	None	Mock	*Giardia* RdRP-AS	*Giardia* Dicer-AS
VSP9B10	99 ± 0.5	98 ± 1.2	90 ± 0.6	18 ± 2.0
VSP1267	0	0.5 ± 0.1	96 ± 0.2	22 ± 0.9
VSPA6	0	0	48 ± 2.3	17 ± 1.3
VSPS1	0	0	62 ± 4.1	36 ± 2.1
VSPS2	0	0	33 ± 1.1	28 ± 3.9
VSPS7	0	0	73 ± 0.3	65 ± 4.4

Percentage of *Giardia* trophozoites expressing a particular VSP as determined by indirect immunofluorescence assays using specific monoclonal antibodies (VSP9B10, monoclonal antibody 9B10; VSP1267, 5C1; VSPA6, 6A7; VSPS1, 1B2; VSPS2, 7B8, VSPS7, 6F8) on WB9B10 trophozoites transfected with the antisense construct of genes encoding *Giardia* RdRP and Dicer, with empty plasmid only (mock transfected) or non-transfected (none), after 5 days in culture. Goat anti-mouse immunoglobulins were used as a control and showed no reaction (not shown). Data represent the mean ± s.d. of three independent experiments performed in duplicate. Results indicate that many different VSPs are expressed in single trophozoites, as judged by addition of each percentage.

encystation[28]. Because CWP genes also comprise a family of homologous genes, it is possible that, in addition to their transcriptional regulation[29], a post-transcriptional process may be affecting CWP mRNA stability.

It is still an open question how *Giardia* changes expression of one VSP for another. Because VSP switching occurs randomly, we favour the hypothesis that variations in general or local concentrations of different VSP transcripts, as occur in higher eukaryotes[30,31], may determine which one will circumvent the silencing system. It is then possible that during cell division each daughter switches to the expression of a different VSP. Alternatively, there may be differences in the level of transcription for each *vsp*. Consistently with this hypothesis, we have detected variations in individual *vsp* expression levels in nuclear run-on experiments performed as in Fig. 1. This sort of variation might depend of the chromatin state in different areas of the genome or the position of each gene within the genome. In this regard, it has been suggested recently that a possible epigenetic mechanism is involved in *Giardia*'s antigenic variation[8].

In summary, a PTGS system (comprising at least RdRP and Dicer) is implicated in the regulation of surface antigen expression in *G. lamblia*. Because a humoral immune response in both experimental and natural *Giardia*-infected hosts coincides with the elimination of the original VSP, a functional role for cells and antibodies in the selection of phenotypic variants and on the course of infection was proposed[32]. Given that parasite protection against specific immune responses relies on switching the expression between immunologically distinct surface proteins[2,3], one way in which hosts can prevent infections with a specific immunological response is by producing antibodies against all surface antigenic determinants. In this report, we have shown that downregulation of components of the RNAi machinery leads to expression of more than one surface antigen in trophozoites. These cells could constitute a fundamental tool for generating a vaccine against this important human pathogen.

METHODS SUMMARY

Parasite culture and cloning. *Giardia* trophozoites were cultured in TYI-33 medium supplemented with adult bovine serum and bovine bile[33]. Continuous cloning of trophozoites was carried out by limiting dilution and selection based on immunofluorescence assays using the corresponding VSP monoclonal antibody. Encystation was carried out as reported[33].

Nuclear run-on and RT–PCR. Nuclear run-on analyses on *Giardia* nuclear extracts were performed as described[34], with some modifications. Total RNA of trophozoites from different isolates was isolated using Trizol (Invitrogen), quantified spectrophotometrically, and treated with RQ1 RNase-Free DNase (Promega). First-strand cDNA was synthesized with Superscript III Reverse Transcriptase (Invitrogen).

Detection and cloning of small VSP RNAs. The detection of small RNA was performed according to ref. 35. [^{32}P]-labelled riboprobes were transcribed *in vitro* by T7 or SP6 RNA polymerase using VSP genes *vsp9B10*, *vsp1267* and *vspH7* cloned into p-GEM T-easy vector (Promega). Labelled RNA was partially hydrolysed by incubation at 60 °C in the presence of $NaHCO_3$ and Na_2CO_3. Isolation and cloning of small RNAs was performed essentially as described previously[36].

Endonuclease activity. Dicer activity was analysed by incubating dsRNA molecules with cytoplasmic extracts of different *Giardia* clones. VSP and control genes cloned into p-GEM T-easy vector (Promega) were transcribed *in vitro* to produce full-length sense [^{32}P]-labelled RNA probes. Pure or mixed VSP transcripts were incubated at 37 °C with *Giardia* extracts for different periods.

Transfection and immunofluorescence assays. The plasmid pTUBPac[37] was modified to introduce sense or antisense *Giardia dicer*, *rdrp* and *ago* coding regions and the influenza haemagglutinin epitope[23]. Transfection of *G. lamblia* trophozoites was done by electroporation as described previously[38]. Indirect immunofluorescence assays were performed on non-encysting trophozoites by using an anti-haemagglutinin monoclonal antibody (Sigma)[23]. For functional analysis of *Giardia dicer*, *rdrp* and *ago*, gene knockdowns were confirmed by RT–PCR using gene-specific primers and by northern blotting.

Received 20 August; accepted 6 October 2008.

1. Deitsch, K. W., Moxon, E. R. & Wellems, T. E. Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol. Mol. Biol. Rev.* **61,** 281–293 (1997).
2. Adam, R. D. Biology of *Giardia lamblia. Clin. Microbiol. Rev.* **14,** 447–475 (2001).
3. Nash, T. E. Antigenic variation in *Giardia lamblia* and the host's immune response. *Phil. Trans. R. Soc. Lond. B* **352,** 1369–1375 (1997).
4. Morrison, H. G. *et al.* Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia. Science* **317,** 1921–1926 (2007).
5. Adam, R. D. *et al.* Antigenic variation of a cysteine-rich protein in *Giardia lamblia. J. Exp. Med.* **167,** 109–118 (1988).
6. Nash, T. E., Mowatt, M. R. & Conrad, J. T. Variant-specific surface protein switching in *Giardia lamblia. Infect. Immun.* **69,** 1922–1923 (2001).
7. Nash, T. E., Alling, D. W., Merritt, J. W. Jr & Conrad, J. T. Frequency of variant antigens in *Giardia lamblia. Exp. Parasitol.* **71,** 415–421 (1990).
8. Kulakova, L. S., Conrad, J. T. & Nash, T. E. Epigenetic mechanisms are involved in the control of *Giardia lamblia* antigenic variation. *Mol. Microbiol.* **61,** 1533–1542 (2006).
9. Cogoni, C. & Macino, G. Homology-dependent gene silencing in plants and fungi: a number of variations on the same theme. *Curr. Opin. Microbiol.* **2,** 657–662 (1999).
10. Fagard, M., Boutet, S., Morel, J. B., Bellini, C. & Vaucheret, H. AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl Acad. Sci. USA* **97,** 11650–11654 (2000).
11. Elmendorf, H. G., Singer, S. M. & Nash, T. E. The abundance of sterile transcripts in *Giardia lamblia. Nucleic Acids Res.* **29,** 4674–4683 (2001).
12. Hutvágner, G. & Zamore, P. D. RNAi: nature abhors a double-strand. *Curr. Opin. Genet. Dev.* **12,** 225–232 (2002).
13. Pak, J. F. *et al.* Distinct populations of primary and secondary effectors during RNAi in *C. elegans. Science* **315,** 241–244 (2007).
14. Dougherty, W. G. P. Transgenes and gene suppression: telling us something new? *Curr. Opin. Cell Biol.* **7,** 399–405 (1995).
15. Zamore, P. D. Ancient pathways programmed by small RNAs. *Science* **296,** 1265–1269 (2002).
16. White, T. C. & Wang, C. C. RNA dependent RNA polymerase activity associated with the double-stranded RNA virus of *Giardia lamblia. Nucleic Acids Res.* **18,** 553–559 (1990).
17. Green-Willms, N. S., Fox, T. D. & Costanzo, M. C. Functional interactions between yeast mitochondrial ribosomes and mRNA 5′ untranslated leaders. *Mol. Cell. Biol.* **18,** 1826–1834 (1998).
18. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).
19. Macrae, I. J. *et al.* Structural basis for double-stranded RNA processing by Dicer. *Science* **311,** 195–198 (2006).
20. Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107,** 309–321 (2001).
21. Höck, J. M. The Argonaute protein family. *Genome Biol.* **9,** 210 (2008).
22. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila. Mol. Cell* **9,** 315–327 (2002).
23. Touz, M. C., Gottig, N., Nash, T. E. & Lujan, H. D. Identification and characterization of a novel secretory granule calcium-binding protein from the early branching eukaryote *Giardia lamblia. J. Biol. Chem.* **277,** 50557–50563 (2002).
24. Elmendorf, H. G. *et al.* Initiator and upstream elements in the α2-tubulin promoter of *Giardia lamblia. Mol. Biochem. Parasitol.* **113,** 157–169 (2001).
25. Ullu, E. L., Lujan, H. D. & Tschudi, C. Small sense and antisense RNAs derived from a telomeric retroposon family in *Giardia intestinalis. Euk. Cell* **4,** 1155–1157 (2005).
26. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453,** 534–543 (2008).
27. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453,** 544–547 (2008).
28. von Allmen, N., Bienz, M., Hemphill, A. & Muller, N. Quantitative assessment of sense and antisense transcripts from genes involved in antigenic variation (*vsp* genes) and encystation (*cwp 1* gene) of *Giardia lamblia* clone GS/M-83–H7. *Parasitology* **130,** 389–396 (2005).
29. Lujan, H. D., Mowatt, M. R. & Nash, T. E. Molecular mechanisms of *Giardia lamblia* differentiation into cysts. *Microbiol. Mol. Biol. Rev.* **61,** 294–304 (1997).
30. Kloc, M. *et al.* RNA localization and germ cell determination in *Xenopus. Int. Rev. Cytol.* **203,** 63–91 (2001).
31. Jambhekar, A. D. *et al. Cis*-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA* **13,** 625–642 (2007).
32. Nash, T. E., Aggarwal, A., Adam, R. D., Conrad, J. T. & Merritt, J. W. Jr. Antigenic variation in *Giardia lamblia. J. Immunol.* **141,** 636–641 (1988).
33. Lujan, H. D., Mowatt, M. R., Conrad, J. T., Bowers, B. & Nash, T. E. Identification of a novel *Giardia lamblia* cyst wall protein with leucine-rich repeats. *J. Biol. Chem.* **270,** 29307–29313 (1995).
34. Yee, J., Mowatt, M. R., Dennis, P. P. & Nash, T. E. Transcriptional analysis of the glutamate dehydrogenase gene in the primitive eukaryote, *Giardia lamblia. J. Biol. Chem.* **275,** 11432–11439 (2000).
35. Hutvagner, G., Mlynarova, L. & Nap, J. P. Detailed characterization of the posttranscriptional gene-silencing-related small RNA in a GUS gene-silenced tobacco. *RNA* **6,** 1445–1454 (2000).
36. Ngo, H., Tschudi, C., Gull, K. & Ullu, E. Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei. Proc. Natl Acad. Sci. USA* **95,** 14687–14692 (1998).
37. Singer, S. M., Yee, J. & Nash, T. E. Episomal and integrated maintenance of foreign DNA in *Giardia lamblia. Mol. Biochem. Parasitol.* **92,** 59–69 (1998).
38. Yee, J. & Nash, T. E. Transient transfection and expression of firefly luciferase in *Giardia lamblia. Proc. Natl Acad. Sci. USA* **92,** 5615–5619 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Gottig and M. E. Alvarez for technical support. This work was supported by grants from the Agencia Nacional para la Promoción de la Ciencia y la Tecnología (ANPCYT), the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), the Universidad Católica de Córdoba (UCC), the Howard Hughes Medical Institute (HHMI), and the European Union CONTENT project. H.D.L. is an HHMI International Research Scholar and a Member of the Scientific Investigator's Career of the CONICET.

Author Contributions C.G.P. knocked down the expression of AGO and VSP9B10, expressed VSPH7 in WB strain trophozoites, performed confocal immunofluorescence assays, northern blots and quantitative RT–PCRs, and cloned and sequenced small RNAs; I.S. knocked down the expression of Dicer and RdRP, performed nuclear run-on and Dicer activity experiments, and cloned and sequenced RdRP, Dicer and VSP genes; R.Q. performed immunofluorescence assays, quantitative RT–PCRs and flow cytometry experiments; P.G.C. performed DNA methylation experiments; and F.D.R., E.V.E. and A.S. generated different monoclonal antibodies and performed immunofluorescence and immunoblotting assays. C.G.P., I.S. and H.D.L wrote this manuscript. H.D.L. conceived and coordinated the project. All authors discussed the results and commented on the manuscript.

Author Information The GenBank accession numbers for *Giardia* Dicer, RdRP and AGO are AY142144, AF293414 and AY142143, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.D.L. (hlujan@ucc.edu.ar).

Vol 456 | 11 December 2008 | doi:10.1038/nature07513

ARTICLES

Negative feedback that improves information transmission in yeast signalling

Richard C. Yu[1], C. Gustavo Pesce[1], Alejandro Colman-Lerner[1]†, Larry Lok[1]†, David Pincus[1]†, Eduard Serra[1]†, Mark Holl[2]†, Kirsten Benjamin[1]†, Andrew Gordon[1]† & Roger Brent[1]

Haploid *Saccharomyces cerevisiae* yeast cells use a prototypic cell signalling system to transmit information about the extracellular concentration of mating pheromone secreted by potential mating partners. The ability of cells to respond distinguishably to different pheromone concentrations depends on how much information about pheromone concentration the system can transmit. Here we show that the mitogen-activated protein kinase Fus3 mediates fast-acting negative feedback that adjusts the dose response of the downstream system response to match the dose response of receptor-ligand binding. This 'dose–response alignment', defined by a linear relationship between receptor occupancy and downstream response, can improve the fidelity of information transmission by making downstream responses corresponding to different receptor occupancies more distinguishable and reducing amplification of stochastic noise during signal transmission. We also show that one target of the feedback is a previously uncharacterized signal-promoting function of the regulator of G-protein signalling protein Sst2. Our work suggests that negative feedback is a general mechanism used in signalling systems to align dose responses and thereby increase the fidelity of information transmission.

Cells use signalling systems to sense and transmit information about extracellular conditions. Haploid *Saccharomyces cerevisiae* yeast cells use a prototypic, G-protein-coupled-receptor/mitogen-activated protein kinase (MAPK) cascade signalling system, the pheromone response system[1], to sense and transmit information about the concentration of mating pheromone secreted by cells of the opposite mating type (Fig. 1). The more information about pheromone concentration the system can transmit, the better a cell can distinguish between different pheromone concentrations, an essential ability for proper partner choice and mating. For example, a yeast cell ringed by potential mating partners strongly prefers to mate with partners producing the most pheromone[2]. Partner choice involves two processes that require sensing of pheromone concentration. First, a cell grows up the pheromone concentration gradient[3], a process that probably depends on measurement of precise differences in pheromone concentration at different points on the cell surface. Second, after contacting its partner and forming a prezygote, a cell preferentially completes fusion and forms a diploid with a partner that produces high amounts of pheromone[4]. These experiments indicate that it is important for cells to distinguish among different pheromone concentrations at multiple steps during the mating process.

Previous work suggested that optimal transmission of information about pheromone concentration depends on both distinguishable receptor occupancies and distinguishable downstream system responses. Differences in receptor occupancy are clearly important for mating partner choice and discrimination; for example, in the presence of pheromone at a concentration that saturates the receptor, cells lose the ability to discriminate between high-concentration pheromone-secreting partners and low-concentration pheromone-secreting partners[2]. However, distinguishable receptor occupancies are not sufficient for partner discrimination, as hypersensitive cells, in the presence of pheromone at a concentration that does not saturate the receptor but does saturate downstream responses, also lose the ability to discriminate between partners secreting different levels of pheromone[2]. One complementary study of mating projection orientation in spatial gradients of pheromone showed that hypersensitive cells do not orient their mating projections as precisely as wild-type cells, perhaps because in these cells downstream responses were saturated at most points in the gradient[3]. However, even in pheromone gradient concentrations 100-fold lower, at which downstream responses should not have been saturated, hypersensitive cells still oriented their mating projections less precisely than did wild-type cells in higher concentration gradients[3]. These observations suggest that hypersensitive cells are inherently less able to respond distinguishably to different pheromone concentrations (that is, transmit less information about pheromone concentration), even when they are responding to pheromone concentrations that saturate neither receptor nor downstream responses.

Distinguishable responses and dose–response alignment

One characteristic of wild-type cells that we[5] and others[6] have previously found is that, despite the large number of intermediate signalling events in the system, the dose–response curve of receptor occupancy closely aligns with dose–response curves of downstream system responses. For example, we observe 'dose–response alignment' between receptor occupancy and the accumulated amount

[1]Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA. [2]Microscale Life Sciences Center, University of Washington, Seattle, Washington 98195, USA. †Present addresses: Instituto de Fisiología, Biología Molecular y Neurociencias, CONICET and Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina (A.C.-L.); Synopsys, Mountain View, California 94043, USA (L.L.); Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158, USA (D.P.); Centre de Genètica Mèdica i Molecular, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona 08907, Spain (E.S.); Biodesign Institute, Arizona State University, Tempe, Arizona 85387, USA (M.H.); Amyris Biotechnologies, Emeryville, California 94608, USA (K.B.); Physics Department, Brookhaven National Laboratory, Upton, New York 11973, USA (A.G.).

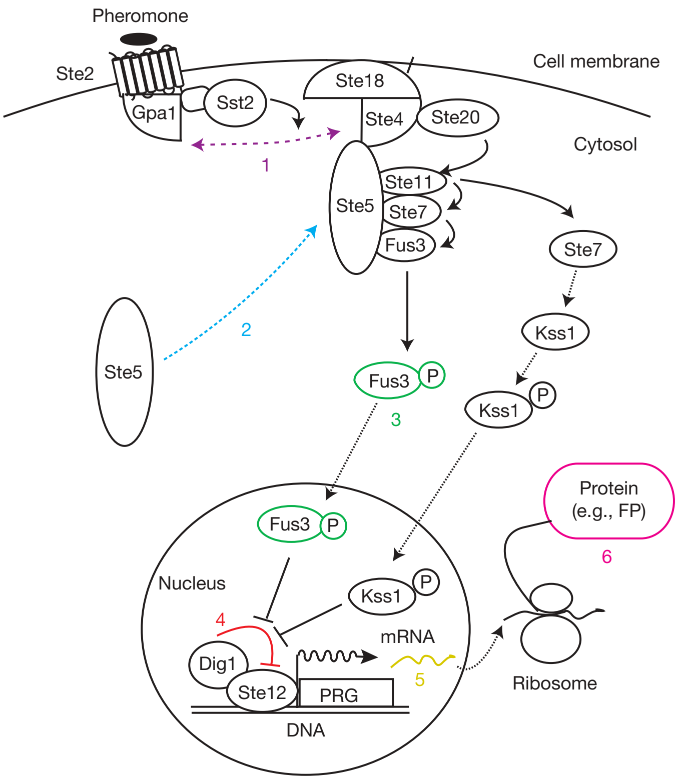


Figure 1 | The pheromone response system. Proteins are indicated by labelled ovals, translocation by dotted arrows, protein activation by arrows, inhibition by T-bar arrows, and protein association by double-headed dashed arrows. Pheromone binding by receptor Ste2 causes dissociation of the heterotrimeric G-protein (1) into Gα subunit (Gpa1) and the Gβγ dimer (Ste4–Ste18). GTP-activating protein (GAP) function of the RGS protein Sst2 promotes re-association of Gpa1 with Ste4–Ste18. On dissociation of the G protein, Ste4 helps recruit the MAPK scaffold Ste5 to the membrane (2). Ste5 recruitment activates the MAPK cascade, in which Ste20, Ste11, Ste7 and the MAP kinases Fus3 and Kss1 phosphorylate one another in sequence. Phosphorylated Fus3 (3) translocates to the nucleus and phosphorylates Dig1 and Ste12, eliminating Dig1 repression of Ste12, a transcriptional activator (4). Ste12 activates transcription of pheromone-responsive genes (PRGs) (5, 6).

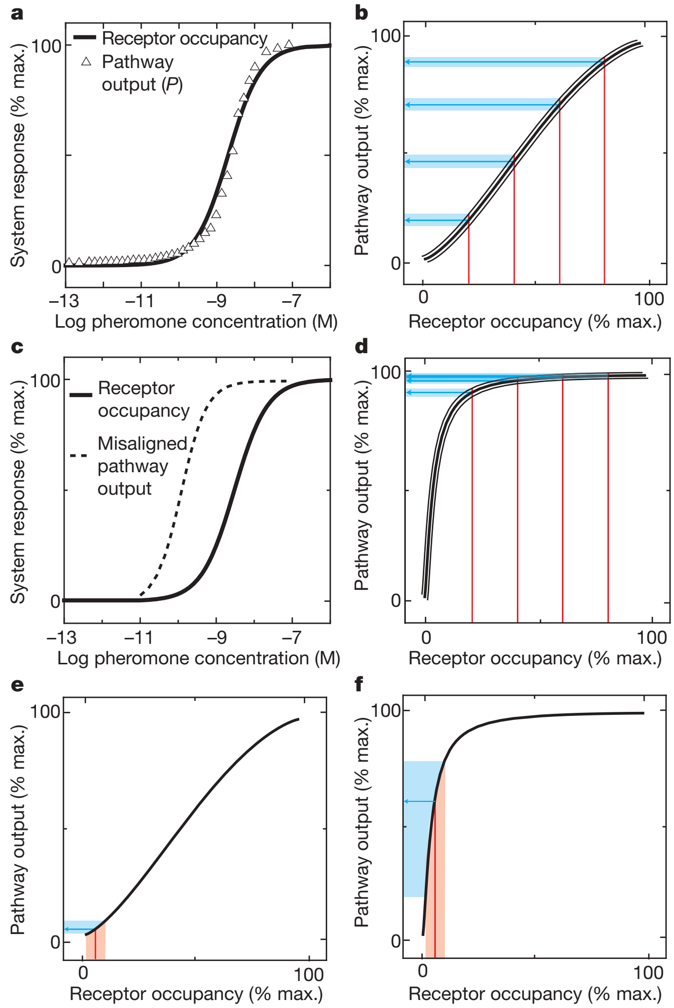


Figure 2 | Dose–response alignment makes responses more distinguishable. **a**, Dose responses of receptor occupancy (calculated from reported receptor–pheromone binding affinity measurements[45,46]) and reporter gene expression output corrected for known sources of cell-to-cell variation (pathway output P[5]) align closely. **b**, The relationship between receptor occupancy and downstream response (from **a**) is essentially linear. Evenly distributed receptor occupancies (20%, 40%, 60% and 80%, red vertical lines) correspond to evenly spaced downstream responses (blue horizontal lines). **c**, Example of dose–response misalignment, in which the downstream output is 20-fold more sensitive than that in **a** (that is, the EC_{50} is reduced 20-fold). **d**, Dose–response misalignment makes transfer function nonlinear, which compresses the downstream responses (blue horizontal lines) corresponding to the majority of receptor occupancies (red vertical lines), making downstream responses less distinguishable. **e**, **f**, Dose–response misalignment results in noise amplification. Receptor occupancy (red vertical line) with noise (pink spread) yields downstream responses (horizontal blue lines) with associated noise (blue spread). In the system with dose–response alignment (**e**), linear transfer function yields less relative noise in the downstream response than in the system with misaligned dose responses and nonlinear transfer function (**f**).

of pheromone-activated Ste12 (cumulative pathway subsystem output P since system induction, where P is operationally defined by reporter-gene expression corrected for inherent cell-to-cell differences in the ability to express proteins[5]; Fig. 2a). Notably, dose–response alignment is commonly observed in many mammalian cell signalling systems, including the insulin[7], acetylcholine[8], thyroid stimulating hormone[9], angiotensin II[10], and epidermal growth factor[11,12] response systems. Researchers in the past have often regarded alignment of curves for ligand binding by a candidate receptor and downstream response as evidence that the putative receptor was in fact the molecule that bound ligand and caused the cellular responses[13–15]. However, to our knowledge, researchers have investigated neither the implications of dose–response alignment for yeast pheromone response nor its general consequences for the function of cell signalling systems.

We realized that dose–response alignment improves information transmission in two ways. First, dose–response alignment describes a linear relationship between receptor occupancy and downstream response; consequently, the entire range of receptor occupancies evenly corresponds to the entire range of possible responses (Fig. 2b). By contrast, even a modest dose–response misalignment, such as a 20-fold shift in the effector concentration for half-maximum response (EC_{50}) of downstream response (Fig. 2c), compresses the downstream responses corresponding to a wide range of receptor occupancies into a narrow range (Fig. 2d). Second, dose–response alignment minimizes the amplification of upstream noise (Fig. 2e). Previous analysis of noise propagation in a synthetic gene circuit revealed analogous amplification of upstream noise in a system with misaligned dose responses[16]. This reasoning suggested to us that cell signalling systems with misaligned dose responses inherently transmit information with lower fidelity, even if downstream responses are not saturated, an idea consistent with previous observations that hypersensitive cells oriented mating projections less precisely in gradients than wild-type cells even at concentrations that did not saturate downstream responses[3]. We hypothesized that dose–response alignment might be indicative of a system that can transmit large

amounts of information, and therefore we sought to understand better the underlying molecular mechanisms required for dose–response alignment and the linear relationship between upstream and downstream response that it defines.

Work in electronics in the 1930s established that 'proportional' negative feedback in electrical circuits, where a constant fraction of the output is subtracted from the input, can bring about a linear input–output voltage relationship[17]. Biologists have also shown that negative feedback can make input–output relationships more linear in biological systems: for example, it was shown that in a MAPK/protein kinase C (PKC)-mediated signalling system, increasing the amount of a MAPK-activated phosphatase that inactivates the MAPK made the average output response more linearly related to (that is, proportional to) the input[18]. Research in both biology and engineering[17,19–22] has also suggested or shown that negative feedback can increase the signal-to-noise ratio in system output and decrease the sensitivity of output to variation in properties of system components (see Supplementary Information 8 for further discussion). These observations suggested to us that negative feedback might mediate dose–response alignment and improve information transmission in the yeast pheromone response system.

Previous work showed that the pheromone response system quickly establishes dose–response alignment: the accumulation of reporter gene expression in cells increased linearly from 15 min to 3 h after stimulation (see Fig. 2 in ref. 5), and at all times the normalized dose response of downstream output aligned with the receptor-ligand-binding curve. These facts suggested that the molecular mechanisms that bring about and stabilize dose–response alignment occur in the first 15 min of pheromone stimulation. However, no quantitative measurements of system activities in this time frame existed to indicate the action of negative feedback to align dose responses. We therefore developed tools to measure the early dynamics of molecular events that the system uses to operate before, during and after establishment of dose–response alignment.

Initial system dynamics indicate negative feedback

We developed reporters and methods to measure real-time signal transmission in single cells, at the membrane and in the nucleus, and supplemented these data with biochemical measurements. We then measured system outputs (that is, system activities at different stages in the signalling pathway, see Fig. 1) after stimulating cells with 100 nM pheromone, a concentration that produces maximal downstream transcription reporter response (Fig. 2a).

Two membrane-proximal system outputs, G-protein activation and Ste5 recruitment to the membrane, peaked and declined rapidly. To follow G-protein activation in single cells over time, we measured loss of fluorescence resonance energy transfer (FRET) between cyan fluorescent protein (CFP)-tagged Gpa1 and yellow fluorescent protein (YFP)-tagged Ste18 by image cytometry[23] (Supplementary Information 2) in a derivative of a strain developed previously[6]. Loss of G-protein FRET rapidly peaked in the first minute and declined (Fig. 3a and Supplementary Fig. 2), consistent with lower time resolution, single-time-point population measurements in an earlier study[6].

We then measured, also in single cells and at subminute intervals, a subsequent membrane-proximal signalling event, the recruitment of Ste5 to the membrane. To do this, we measured the redistribution of YFP–Ste5 from the nucleus and cytosol to the membrane (Supplementary Information 2). Membrane recruitment of Ste5 was rapid (Fig. 3b). Within 5 s of stimulation with a high concentration of pheromone, individual cells showed an increase in yellow fluorescence at the cell membrane, and a corresponding depletion of fluorescence from the cell interior—no change in fluorescence was observed in unstimulated cells or cells with unlabelled Ste5 (Supplementary Fig. 3). Compared with unstimulated cells, average membrane recruitment reached near-maximal values within seconds and peaked by 20 s, before declining towards a plateau in later minutes, similar to the dynamics of G-protein loss of FRET.

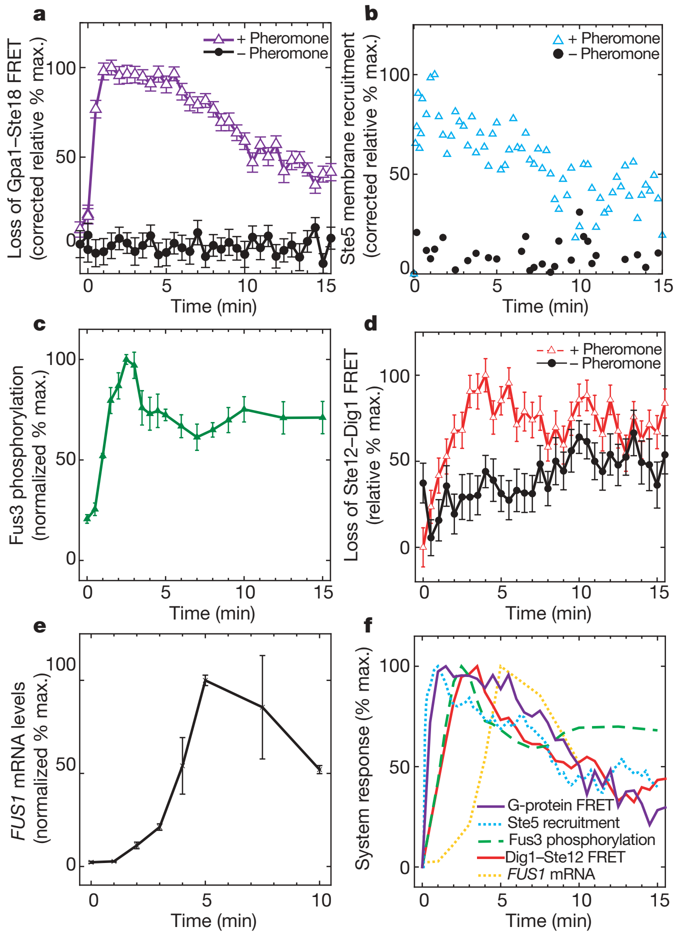


Figure 3 | Initial system dynamics indicate negative feedback. a, Loss of G-protein FRET. Corrected median (± standard error) loss of Gpa1–Ste18 FRET values (relative to maximum change measured in pheromone-stimulated cells; see Supplementary Fig. 2) in RY2062b cells stimulated with pheromone (purple triangles; $n = 262$) quickly peaked and declined to a plateau relative to unstimulated cells (black circles; $n = 143$). **b**, YFP–Ste5 recruitment. Corrected median (± standard error) YFP–Ste5 membrane recruitment (relative to maximum change measured in pheromone-stimulated cells; see Supplementary Fig. 3e) in RY2013 cells stimulated with pheromone stimulation (cyan triangles, $n = 361$) quickly peaked and declined to a plateau compared to unstimulated cells (black circles, $n = 223$). **c**, Fus3 activity. Mean ratios (± standard error, $n = 3$–5) of activated (phospho-Y180 and phospho-T182) Fus3 to total Fus3, normalized to the peak measured ratio (see Supplementary Fig. 4 for representative immunoblot images). Fus3 activity levels peaked 2.5 min after pheromone stimulation and declined to a plateau within 5 min of stimulation. Total Fus3 levels, compared to levels of non-pheromone-regulated proteins GAPDH and PGK1, remained constant over this time period (data not shown). New protein synthesis is not required for the observed peak and decline (Supplementary Fig. 4c). **d**, Loss of Dig1–Ste12 FRET. Median (± standard error) loss of Dig1–Ste12 FRET (scaled to minimum and maximum values measured in pheromone-stimulated cells; for raw values, see Supplementary Fig. 7) in RY1130b cells peaked about 3 min after pheromone stimulation, and then declined to a plateau (red triangles, $n = 246$) relative to unstimulated cells (black circles, $n = 138$). **e**, *FUS1* mRNA. Average ratio (high and low values indicated) of *FUS1* mRNA probe band intensity to loading control (*ACT1* mRNA probe band intensity) after pheromone stimulation ($n = 2$) (see Supplementary Fig. 8 for raw image). **f**, Composite timing plot shows persistent peak-and-decline towards a plateau for all system responses, suggesting action of negative feedback. YFP–Ste5 recruitment and Dig1–Ste12 FRET (from **b** and **d**) were smoothed using a moving window of five data points.

We then assessed intermediate system output farther downstream by measuring the dynamics of MAPK activation. Using quantitative immunoblotting, we measured phosphorylation of Fus3 residues

Thr 180 and Tyr 182, which is required for Fus3 activity and pheromone response[24]. The amount of phosphorylated Fus3 relative to total Fus3 increased rapidly, reaching a maximum in 2.5 min before dropping to a plateau level in approximately 5–7 min (Fig. 3c and Supplementary Fig. 4a, b).

We then measured nuclear MAPK activity dynamics in single cells over time. To do this we developed a FRET reporter to measure pheromone-induced changes in the association between the transcription factor Ste12 and one of its inhibitors, Dig1 (ref. 25). We deleted native *STE12* and *DIG1* genes and chromosomally integrated versions of these proteins fused to CFP and YFP, respectively (see Supplementary Information 5). We then measured changes in FRET between CFP and YFP[26] in the nucleus by image cytometry[23] (see Supplementary Fig. 6 and Supplementary Information 2). Pheromone-induced loss of FRET did not require new protein translation (Supplementary Fig. 7b), but did require both Ste5 (Fig S7c) and MAPK activity (Supplementary Fig. 7d), consistent with the interpretation that loss of FRET directly reports pheromone-induced, MAPK-mediated de-repression of Ste12.

Loss of Dig1–Ste12 FRET quickly peaked around 3 min after stimulation (Fig. 3d), and the overall signal dynamics were very similar to those of Fus3 phosphorylation (compare Fig. 3c to Fig. 3d; see also Fig. 3f). This fast signal transfer from Fus3 activation to Ste12 de-repression is consistent with the idea that Fus3 moves quickly in and out of the nucleus, as shown in studies of changes in Fus3 localization by fluorescence recovery after photobleaching (FRAP)[27]. We confirmed that the timing of Ste12 de-repression measured by loss of Ste12–Dig1 FRET was consistent with the dynamics of pheromone-induced mRNA transcription. Using RNase protection assays, we quantified *FUS1* mRNA levels. *FUS1* mRNA levels peaked at 5 min after pheromone stimulation before declining (Fig. 3e and Supplementary Fig. 8). The maximum rate of increase in mRNA occurred between 3 and 5 min, consistent with the time of maximum loss of Ste12–Dig1 FRET.

All measurements of signal-relaying events showed a consistent pattern of rapid peak and decline towards a plateau after pheromone stimulation (Fig. 3f), which suggested the action of one or more fast-acting negative feedbacks that might modulate the dose-dependence of the signal to achieve dose–response alignment.

Fus3 mediates negative feedback

A number of previous studies suggested that the MAPKs Fus3 and Kss1 might mediate rapid negative feedback. Our previous study of regulated cell-to-cell variation in system output revealed a Fus3-dependent reduction in variation, suggesting an autoregulatory negative feedback mediated by Fus3 (ref. 5). One study showed that levels of phosphorylated Fus3 were higher in cells bearing a kinase-dead mutant version of Fus3 (ref. 24). Another study showed that Ste5(T287A) mutant cells, in which the Ste5 protein is mutated at the site of a threonine residue phosphorylated by Fus3 on Ste5-derived peptides *in vitro*, showed increased reporter expression[28], albeit with no change in the EC_{50} of the dose response. Another study found that glutamate substitutions at candidate sites of Fus3 phosphorylation on Ste7 decreased on-scaffold Fus3 activation and mating efficiency[29]. Additionally, phosphoproteomic studies of pheromone response system proteins[30] have uncovered numerous sites of phosphorylation on system proteins, and many of these phosphor-

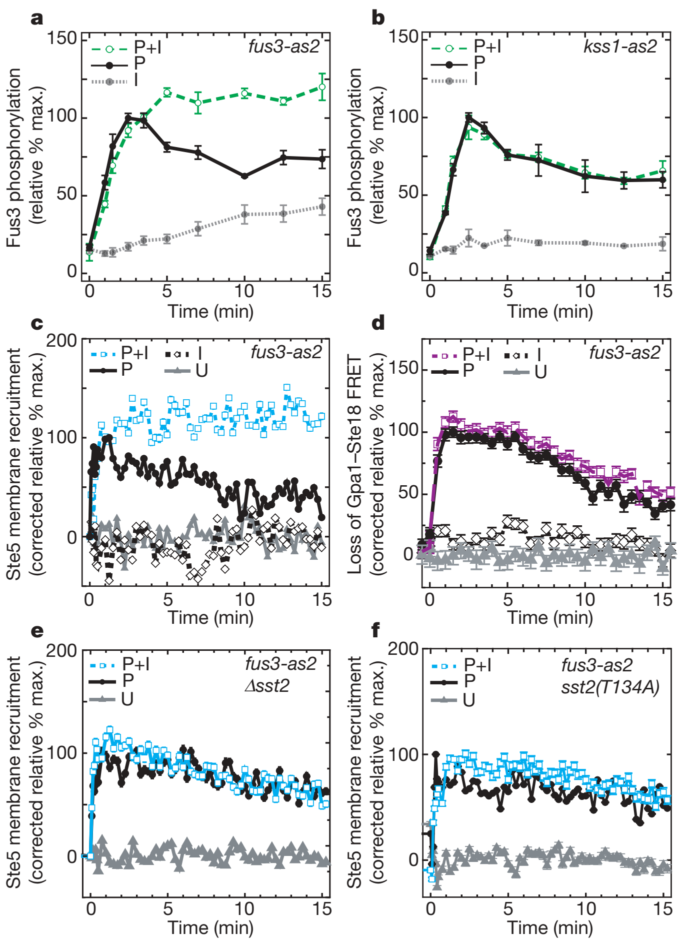


Figure 4 | Fus3 mediates negative feedback. Values are scaled to peak signal measured in cells stimulated with only pheromone. Error bars indicate ± standard error for all panels: P, stimulated with 100 nM pheromone; P+I, stimulated with 100 nM pheromone plus 10 μM 1-NM-PP1; I, 10 μM 1-NM-PP1; U, untreated. **a**, Fus3 mediates negative feedback. In *fus3-as2* cells (RY1134b), mean ($n = 4$) Fus3 phosphorylation peaked and declined, as in *FUS3* cells (Fig. 3b), after pheromone stimulation (black circles), but did not decline when we stimulated cells simultaneously with Fus3-as2 inhibitor (green circles). Treating cells with only inhibitor (grey circles) caused the signal to rise slowly, indicating that cells actively regulate basal signal level. **b**, Kss1 does not mediate negative feedback. In *kss1-as2* cells (RY1133b), mean ($n = 4$) Fus3 phosphorylation in pheromone-stimulated cells without (black circles) or with (green circles) simultaneous treatment with Kss1-as2 inhibitor were identical. Treating cells with only inhibitor caused no significant increase in Fus3 phosphorylation (grey circles). **c**, Fus3-mediated feedback acts on or upstream of Ste5 membrane recruitment. In *fus3-as2* cells (RY2013), median YFP–Ste5 membrane recruitment peaked and declined after pheromone stimulation (filled circles; $n = 361$), but did not decline after simultaneous treatment with Fus3-as2 inhibitor (blue squares; $n = 196$). There was no relative Ste5 recruitment in cells treated with inhibitor alone (open diamonds; $n = 134$) or in completely untreated cells (grey triangles; $n = 220$). The small increase in Fus3 phosphorylation measured in cells treated with inhibitor only (black circles in **a**) suggests that additional Fus3-independent mechanisms maintain low basal levels of Ste5 recruitment. **d**, Fus3-mediated negative feedback acts downstream of G-protein dissociation. In *fus3-as2* cells (RY2062b, derived from TMY101 (ref. 6)), median Gpa1–Ste18 loss of FRET peaked and declined in pheromone-stimulated cells (filled circles; $n = 262$) with the same dynamics as in pheromone-stimulated cells simultaneously treated with Fus3-as2 inhibitor (purple squares; $n = 263$). Unstimulated cells in the presence (open diamonds; $n = 229$) or absence (grey triangles; $n = 143$) of inhibitor showed no loss of Gpa1–Ste18 FRET. **e**, One target of Fus3-mediated negative feedback is a novel Sst2-dependent increase in YFP–Ste5 recruitment. Median YFP–Ste5 membrane recruitment in pheromone-stimulated *fus3-as2 Δsst2* cells (RY2024) peaked and declined both in the absence (black circles; $n = 188$) or presence (cyan squares; $n = 300$) of Fus3-as2 inhibitor, similar to *SST2* cells with active Fus3 (black circles, panel **c**). Grey triangles indicate unstimulated cells ($n = 268$). **f**, Mutation of predicted Fus3/MAPK phosphorylation site in Sst2 DEP1 domain eliminated Sst2 promotion of YFP–Ste5 recruitment. Median (± standard error) YFP–Ste5 membrane recruitment in pheromone-stimulated *sst2(T134A)* (RY2077) cells peaked and declined both in the absence (filled circles; $n = 309$) and presence (cyan squares; $n = 321$) of inhibitor, similar to *Δsst2* cells (panel **e**). Grey triangles indicate unstimulated cells ($n = 334$).

ylation sites lie in consensus MAPK target sequences (R. Maxwell and O. Resnekov, personal communication). Finally, work with one pheromone response system model showed that feedbacks, including one from Fus3, increased the alignment of upstream and downstream dose responses[31]. We therefore proposed that the signal decline at different measurement points depends on non-translational, fast-acting negative feedbacks mediated by Fus3 or Kss1.

To test whether Fus3 or Kss1 is a source of negative feedback on system activity, we compared the baseline system response, at system points up to and including Fus3 phosphorylation, with system response after selective inhibition of either Fus3 or Kss1 kinase activity. To do this, we first modified reporter strains by replacing either *FUS3* or *KSS1* with the corresponding purine-analogue-sensitive allele[32]. We did this by changing the 'gatekeeper' residue in each kinase's ATP binding pocket (Q93 in Fus3; N94 in Kss1) to an alanine. The mutant *fus3-as2* and *kss1-as2* kinases were active, as measured by fluorescent protein reporter gene output (Supplementary Fig. 9a), and 10 μM 4-amino-1-(tert-butyl)-3-(1′-naphthylmethyl)-pyrazolo[3,4-*d*]pyrimidine (1-NM-PP1), a cell-permeable adenosine analogue, inhibited the activity of mutant kinases without inhibiting wild-type kinases (Supplementary Fig. 9b). We then quantified Fus3 phosphorylation by quantitative immunoblotting after stimulation with pheromone, either with or without simultaneous inhibition with 1-NM-PP1 (Fig. 4a). Fus3 phosphorylation levels did not peak and decline to a plateau when we inhibited Fus3-as2, but, rather, remained high, near peak levels. By contrast, when we inhibited Kss1-as2, Fus3 phosphorylation levels were unaffected (Fig. 4b). These results indicated that Fus3 kinase activity mediated one or more negative feedbacks in this system.

We then studied where in the system the Fus3-dependent feedback acted to diminish signal amplitude. The decrease in G-protein FRET within 30 s of stimulation depends on Sst2 (ref. 6). This finding suggested that the Fus3-dependent negative feedback might upregulate the GTPase-activating protein (GAP) function of Sst2, which would increase G-protein re-association and decrease downstream signal. We tested whether Fus3-as2 inhibition affected the observed decline in both G-protein dissociation and Ste5 recruitment. Inhibition of Fus3 activity eliminated the decline in Ste5 translocation (Fig. 4c), but surprisingly had no effect on the decline in G-protein dissociation in a G-protein FRET reporter strain carrying *fus3-as2* (Fig. 4d and Supplementary Fig. 10). These results indicated that Fus3-mediated negative feedback acted downstream of mechanisms regulating G-protein association.

To confirm that Fus3 acted downstream of G-protein activation, we measured Ste5 recruitment after deleting *SST2*. We expected deletion of *SST2* to have no effect on Fus3-mediated signal decline, as Sst2 is required for efficient G-protein inactivation and, as we showed above, Fus3-mediated negative feedback does not reduce G-protein dissociation levels. Unexpectedly, when we deleted *SST2*, we completely disrupted Fus3-mediated signal decline; unlike $SST2^+$ cells, inhibition of Fus3 did not cause an increase in Ste5 recruitment (Fig. 4e). Furthermore, Ste5 recruitment (with or without Fus3-mediated feedback) peaked and declined similar to the baseline response of $SST2^+$ cells (compare squares and circles in Fig. 4e with circles in Fig. 4c). This finding showed that signal peak and decline is the default behaviour in the absence of Sst2. Because a sustained non-declining signal is only evident in $SST2^+$ cells in the presence of Fus3-as2 inhibitor, these results also indicate that Sst2 promotes Ste5 membrane recruitment, a hitherto unknown function of the regulator of G-protein signalling (RGS) protein family, and that Fus3 negatively regulates this novel signal-promoting function (Fig. 5a).

We then investigated which portions of the Sst2 protein might be involved in promoting Ste5 membrane recruitment. During analysis of Sst2 point mutants, we found that Ste5 recruitment in a *fus3-as2* strain that carried *sst2(T134A)* instead of wild-type Sst2 peaked and declined in the presence and absence of Fus3 inhibitor (Fig. 4f), just as observed in Δ*sst2* cells (Fig. 4e). The pheromone-induced growth

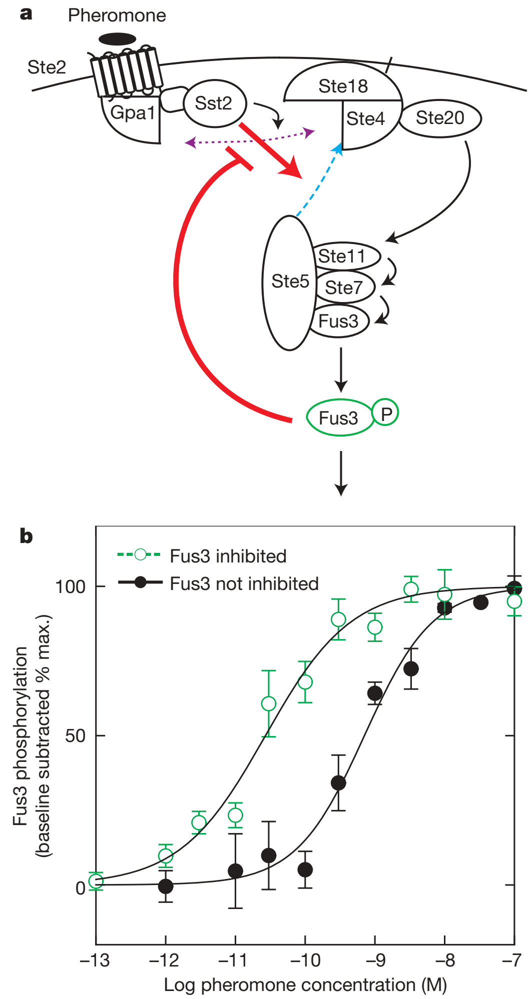


Figure 5 | Dose–response alignment requires Fus3-mediated negative feedback. a, Model of negative feedback regulation of Ste5 membrane recruitment. Sst2 promotes (thick red arrow) Ste5 recruitment to the membrane (blue dashed arrow), and Fus3 negatively regulates this signal promotion (thick T-bar arrow). **b**, Fus3 inhibition disrupts dose–response alignment. In pheromone-stimulated *fus3-as2* (RY2052b) cells, inhibition of Fus3 kinase activity (green circles) reduced the sensitivity (EC_{50}) of the dose response of mean Fus3 phosphorylation ($\pm$ standard error; $n = 3$–4) relative to cells not treated with inhibitor (black filled circles). This measured dose response of Fus3 phosphorylation in uninhibited cells is approximately ten times more sensitive than a previous measurement in a different strain background[47]. Fus3 phosphorylation was measured after 15 min of pheromone stimulation, after the signal reaches the dose-dependent plateau (see Fig. 3c). Black lines show fits to Hill functions. Fus3 inhibition reduced the EC_{50} of the dose response by greater than 20-fold without affecting the gradedness (cooperativity) of the average response (see Supplementary Information for details).

inhibition of *sst2(T134A)* cells reported by halo assays was close to wild-type levels (Supplementary Fig. 11a), and the average number of Sst2(T134A) protein molecules per cell was similar to Sst2 abundance in the parent strain (Supplementary Fig. 11b), suggesting that the T134A mutation disrupts a significant fraction of the Fus3-dependent, signal-promoting function of Sst2 without disrupting the bulk of its signal-reducing GAP activity. T134 lies within the N-terminal DEP domains of Sst2, which are required for localization of Sst2 to the membrane by binding the cytosolic tail of Ste2 (ref. 33). These results indicate that the DEP domains in Sst2 might aid Ste5 membrane recruitment, perhaps by providing additional membrane-proximal interaction surfaces, and suggest that mechanisms that regulate localization of Sst2 to the membrane, such as disruption of Sst2–Ste2 interactions by Yck1/2-mediated phosphorylation after

longer periods of pheromone stimulation[33], might consequently regulate Ste5 membrane recruitment.

Fus3-generated negative feedback aligns dose responses

Finally, we tested whether dose–response alignment between receptor–pheromone binding and downstream activities required Fus3 activity. In principle, Fus3-mediated negative feedback might scale system activity by a dose-independent factor, and therefore cause no shift in the normalized dose–response curve. For example, the Ste5(T287A) mutation increases the magnitude of system output relative to wild-type cells without changing the pheromone concentration yielding half-maximal response (see Fig. 5 in ref. 28). We measured dose responses of Fus3 phosphorylation in a *fus3-as2* strain with and without inhibitor 15 min after pheromone stimulation, the time when the amount of Fus3 phosphorylation had declined to a steady-state level (Fig. 3c). Inhibiting Fus3 kinase activity shifted the dose response of Fus3 activation, lowering the pheromone concentration needed for half-maximal response by 20-fold (Fig. 5b). Moreover, Fus3-mediated negative feedback reduces the dynamic range of the output by only 50% (see Supplementary Information 8.2 and Supplementary Fig. 14). These results showed that Fus3-mediated negative feedback was required for dose–response alignment in the yeast pheromone response system.

Discussion

We found that MAPK Fus3 mediates rapid negative feedback that aligns the dose responses of upstream and downstream system activities in the pheromone response system. We propose that dose–response alignment improves information transmission through this and other signalling systems. Furthermore, we found that Fus3 negatively regulates a novel signal-promoting function of the RGS protein Sst2. Our results demonstrate that RGS proteins, present in many eukaryotic signalling systems (the human RGS family, for example, contains more than 35 members[34]), can function in signal transduction systems by increasing signal in addition to accelerating G-protein inactivation, possibly (as in the case of pheromone response) by facilitating recruitment of MAPK scaffolds to sites of activity.

The idea that dose–response alignment increases the amount of transmitted information has practical implications for drug discovery and design. For example, consider a drug that increased the sensitivity of cells to a naturally occurring antagonist of cell proliferation, analogous to the downstream dose–response shift we observed on Fus3 inhibition in the pheromone response system (Fig. 5b). Despite increasing the average sensitivity of cells to signals to stop growth, the dose–response misalignment could reduce the amount of transmitted information about the signal. The decrease in transmitted information could increase cell-to-cell variation in response, causing a larger number of cells to fall below a threshold in antagonist response and continue proliferation. It is possible that some existing drugs that allosterically modify G-protein-coupled receptor signalling systems downstream of ligand binding (see Fig. 3 in ref. 35) and those that target mid-system signalling molecules such as PKC[36] and AKT[37] may decrease dose–response alignment and increase response variation, whereas drugs that specifically affect the affinity of receptor–ligand binding (see Fig. 5 in ref. 35) should not.

We propose here that the fidelity with which a cell responds to different input concentrations of a ligand depends on a 'systems-level' quantitative behaviour, dose–response alignment, found in many other cell signalling systems. For biological systems, a deeper understanding of key quantitative behaviours will probably depend on articulating appropriate analytical frameworks and metrics. Information theory[38] defines a framework for quantifying the relationship between system input and output (see Supplementary Information 9 for further discussion), and has enabled researchers to quantify, for example, the amount of information that an axon of a single sensory neuron can transmit[39] and the amount of information about morphogen gradient that a transcription factor can transmit to a downstream effector[40,41]. In much the same way as concepts from classical electromagnetism provide rigorous means to describe and understand the determinants of behaviours of electrical circuits, we expect that concepts from information theory will enable more rigorous and quantitative understanding of how the protein components of complex vertebrate cell signalling systems interact to sense and transmit information into the cell.

METHODS SUMMARY

We constructed yeast strains and plasmids by standard methods[42,43] essentially as described (ref. 5 and Supplementary Information 1). By doctrine, we expressed all reporter constructs from native promoters integrated into the chromosome, and verified that the level of expressed protein was similar to the native level. With the exception of strains used for G-protein FRET experiments, we constructed all strains from the otherwise-isogenic *bar1*$^{-}$ W303a reference parent strain, ACL 379 (ref. 5), by the steps described. We stimulated exponentially growing cells with the indicated concentration of pheromone and/or other reagents (such as the inhibitor 1-NM-PP1) in one of two ways. For image cytometry, we affixed the cells to the bottom of wells in a glass-bottom 96-well plate, as described previously[5] and in Supplementary Information 2.1. Using custom fluidic hardware, we evacuated medium from the well, injected fresh medium containing the indicated concentration of pheromone and/or inhibitor, and proceeded to record images over time. For MAPK phosphorylation, *FUS1* mRNA, and flow cytometry experiments, we stimulated cells by using a micropipette to mix a small volume of pheromone and/or inhibitor into the cell suspension to the final concentration (as indicated, typically 100 nM pheromone and 10 μM 1-NM-PP1). We performed image acquisition essentially as described[5,23], with modifications as detailed in Supplementary Information. For image cytometry, we extracted values for parameters of interest from images using Cell-ID 1.0 (ref. 23). We analysed image and flow cytometric data using Physics Analysis Workstation (see ref. 44) and custom scripts, depending on the type of image, described in the text and in Supplementary Information. Supplementary Information contains further details on plasmids, strains, construction methods, materials and experimental methods.

Received 23 December 2007; accepted 3 October 2008.

1. Dohlman, H. G. & Thorner, J. W. Regulation of G protein-initiated signal transduction in yeast: Paradigms and principles. *Annu. Rev. Biochem.* **70**, 703–754 (2001).
2. Jackson, C. L. & Hartwell, L. H. Courtship in *S. cerevisiae*: Both cell types choose mating partners by responding to the strongest pheromone signal. *Cell* **63**, 1039–1051 (1990).
3. Segall, J. E. Polarization of yeast cells in spatial gradients of alpha mating factor. *Proc. Natl Acad. Sci. USA* **90**, 8332–8336 (1993).
4. Schrick, K., Garvik, B. & Hartwell, L. H. Mating in *Saccharomyces cerevisiae*: The role of the pheromone signal transduction pathway in the chemotropic response to pheromone. *Genetics* **147**, 19–32 (1997).
5. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**, 699–706 (2005).
6. Yi, T. M., Kitano, H. & Simon, M. I. A quantitative characterization of the yeast heterotrimeric G protein cycle. *Proc. Natl Acad. Sci. USA* **100**, 10764–10769 (2003).
7. Cuatrecasas, P. Insulin–receptor interactions in adipose tissue cells: Direct measurement and properties. *Proc. Natl Acad. Sci. USA* **68**, 1264–1268 (1971).
8. Kasai, M. & Changeux, J. P. *In vitro* excitation of purified membrane by cholinergic agonists. *J. Membr. Biol.* **6**, 58–80 (1971).
9. Amir, S. M., Carraway, T. F. Jr, Kohn, L. D. & Winand, R. J. The binding of thyrotropin to isolated bovine thyroid plasma membranes. *J. Biol. Chem.* **248**, 4092–4100 (1973).
10. Lin, S. Y. & Goodfriend, T. L. Angiotensin receptors. *Am. J. Physiol.* **218**, 1319–1328 (1970).
11. Knauer, D. J., Wiley, H. S. & Cunningham, D. D. Relationship between epidermal growth factor receptor occupancy and mitogenic response. Quantitative analysis using a steady state model system. *J. Biol. Chem.* **259**, 5623–5631 (1984).
12. Nagashima, T. *et al.* Quantitative transcriptional control of Erbb receptor signaling undergoes graded to biphasic response for cell differentiation. *J. Biol. Chem.* **282**, 4045–4056 (2007).
13. Simons, S. S. Jr, Oshima, H. & Szapary, D. Higher levels of control: Modulation of steroid hormone-regulated gene transcription. *Mol. Endocrinol.* **6**, 995–1002 (1992).
14. Rousseau, G. G. & Baxter, J. D. Glucocorticoid receptors. *Monogr. Endocrinol.* **12**, 49–77 (1979).
15. Bloom, E. *et al.* Nuclear binding of glucocorticoid receptors: Relations between cytosol binding, activation and the biological response. *J. Steroid Biochem.* **12**, 175–184 (1980).

16. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
17. Black, H. S. Stabilized feed-back amplifiers. *Electr. Eng.* **53**, 114–120 (1934).
18. Bhalla, U. S., Ram, P. T. & Iyengar, R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297**, 1018–1023 (2002).
19. Black, J. W. & Leff, P. Operational models of pharmacological agonism. *Proc. R. Soc. Lond. B* **220**, 141–162 (1983).
20. Savageau, M. A. Comparison of classical and autogenous systems of regulation in inducible operons. *Nature* **252**, 546–549 (1974).
21. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
22. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
23. Gordon, A. *et al.* Single-cell quantification of molecules and rates using open-source microscope-based cytometry. *Nature Methods* **4**, 175–181 (2007).
24. Gartner, A., Nasmyth, K. & Ammerer, G. Signal transduction in *Saccharomyces cerevisiae* requires tyrosine and threonine phosphorylation of Fus3 and Kss1. *Genes Dev.* **6**, 1280–1292 (1992).
25. Tedford, K., Kim, S., Sa, D., Stevens, K. & Tyers, M. Regulation of the mating pheromone and invasive growth responses in yeast by two map kinase substrates. *Curr. Biol.* **7**, 228–238 (1997).
26. Miyawaki, A. & Tsien, R. Y. Monitoring protein conformations and interactions by fluorescence resonance energy transfer between mutants of green fluorescent protein. *Methods Enzymol.* **327**, 472–500 (2000).
27. van Drogen, F., Stucke, V. M., Jorritsma, G. & Peter, M. Map kinase dynamics in response to pheromones in budding yeast. *Nature Cell Biol.* **3**, 1051–1059 (2001).
28. Bhattacharyya, R. P. *et al.* The ste5 scaffold allosterically modulates signaling output of the yeast mating pathway. *Science* **311**, 822–826 (2006).
29. Maleri, S. *et al.* Persistent activation by constitutive Ste7 promotes Kss1-mediated invasive growth but fails to support Fus3-dependent mating in yeast. *Mol. Cell. Biol.* **24**, 9221–9238 (2004).
30. Gruhler, A. *et al.* Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4**, 310–327 (2005).
31. Shao, D., Zheng, W., Qiu, W., Ouyang, Q. & Tang, C. Dynamic studies of scaffold-dependent mating pathway in yeast. *Biophys. J.* **91**, 3986–4001 (2006).
32. Bishop, A. C. *et al.* A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407**, 395–401 (2000).
33. Ballon, D. R. *et al.* DEP-domain-mediated regulation of GPCR signaling responses. *Cell* **126**, 1079–1093 (2006).
34. Heximer, S. P. & Blumer, K. J. RGS proteins: Swiss army knives in seven-transmembrane domain receptor signaling networks. *Sci. STKE* **2007**, pe2 (2007).
35. May, L. T., Leach, K., Sexton, P. M. & Christopoulos, A. Allosteric modulation of G protein-coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* **47**, 1–51 (2007).
36. Budas, G. R., Churchill, E. N. & Mochly-Rosen, D. Cardioprotective mechanisms of PKC isozyme-selective activators and inhibitors in the treatment of ischemia-reperfusion injury. *Pharmacol. Res.* **55**, 523–536 (2007).
37. Lindsley, C. W., Barnett, S. F., Layton, M. E. & Bilodeau, M. T. The PI3K/Akt pathway: Recent progress in the development of ATP-competitive and allosteric Akt kinase inhibitors. *Curr. Cancer Drug Targets* **8**, 7–18 (2008).
38. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
39. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).
40. Tkacik, G., Callan, C. & Bialek, W. Information flow and optimization in transcriptional regulation. *Proc. Natl. Acad. Sci. U.S.A.* **26**, 12265–12270 (2008).
41. Gregor, T., Tank, D. W., Wieschaus, E. F. & Bialek, W. Probing the limits to positional information. *Cell* **130**, 153–164 (2007).
42. Ausubel, F. M. *et al. Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., 1987–2008).
43. Guthrie, C. & Fink, G. R. *Methods in Enzymology, Guide to Yeast Genetics and Molecular Biology* (Academic, 1991).
44. Brun, R., Couet, O., Vandroni, C. & Zanarini, O. PAW Physics Analysis Workstation CERN program library entry q121 (CERN, Geneva, 1989).
45. Jenness, D. D., Burkholder, A. C. & Hartwell, L. H. Binding of alpha-factor pheromone to *Saccharomyces cerevisiae* a cells: Dissociation constant and number of binding sites. *Mol. Cell. Biol.* **6**, 318–320 (1986).
46. Bajaj, A. *et al.* A fluorescent α-factor analogue exhibits multiple steps on binding to its g protein coupled receptor in yeast. *Biochemistry* **43**, 13564–13578 (2004).
47. Andersson, J., Simpson, D. M., Qi, M., Wang, Y. & Elion, E. A. Differential input by Ste5 scaffold and Msg5 phosphatase route a MAPK cascade to multiple outcomes. *EMBO J.* **23**, 2564–2576 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Abola, S. Andrews, A. Arkin, M. Bowen, L. Buck, C. Denby, A. Gann, D. Meldrum, T. Mitchison, C. Pabo, M. Ptashne, M. Reese, O. Resnekov, C. Ryan, M. Snyder, T. Thomson, A. E. Tsong and M. Wilson for discussions and/or comments on the manuscript, and O. Resnekov for help in articulating the requirements for fluidic induction devices. Work, including that of M.H. at the University of Washington, was supported by the Alpha Project at the Center for Quantitative Genome Function, an NIH Center of Excellence in Genomic Science under grant P50 HG02370 from the National Human Genome Research Institute to R.B.

Author Contributions R.C.Y. performed the Ste5 translocation, loss of G-protein and Dig1–Ste12 FRET, MAP kinase phosphorylation, mRNA measurements, and the image-based measurements of system output. C.G.P. and A.G. noted and helped articulate the relationship of negative feedback to dose–response overlap. C.G.P. and R.C.Y. performed flow cytometric measurements. L.L. and A.G. contributed to discussions about mutual information calculations and analysis. R.C.Y. and A.G. performed data analysis on loss of FRET experiments. A.C.-L. and A.G. carried out extensive initial measurements of Ste5 recruitment. K.B. provided unpublished information about protein quantification useful in initial discussions about dose–response alignment. R.C.Y. and D.P. performed numerous measurements verifying protein abundance in other strains. E.S. constructed the inhibitor-sensitive *fus3-as2* and *kss1-as2* alleles. M.H. designed and built fluidic devices used to induce the system in response to articulated requirements. R.B. provided input into project direction, experimental design and interpretation of results. R.C.Y., C.G.P. and R.B. wrote the paper and guarantee the integrity of the results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to ryu@molsci.org and brent@molsci.org.

ARTICLES

The replisome uses mRNA as a primer after colliding with RNA polymerase

Richard T. Pomerantz[1] & Mike O'Donnell[1]

Replication forks are impeded by DNA damage and protein–nucleic acid complexes such as transcribing RNA polymerase. For example, head-on collision of the replisome with RNA polymerase results in replication fork arrest. However, co-directional collision of the replisome with RNA polymerase has little or no effect on fork progression. Here we examine co-directional collisions between a replisome and RNA polymerase *in vitro*. We show that the *Escherichia coli* replisome uses the RNA transcript as a primer to continue leading-strand synthesis after the collision with RNA polymerase that is displaced from the DNA. This action results in a discontinuity in the leading strand, yet the replisome remains intact and bound to DNA during the entire process. These findings underscore the notable plasticity by which the replisome operates to circumvent obstacles in its path and may explain why the leading strand is synthesized discontinuously *in vivo*.

DNA damage and high affinity protein–nucleic acid complexes, such as transcribing RNA polymerase (RNAP), act as impediments to bacterial and eukaryotic replication forks[1–5]. Arrest of the replication machinery can lead to mutagenesis and cell death. Thus, several pathways have evolved to repair and restart various types of collapsed replication forks. Mechanisms that facilitate replication past sites of DNA damage, such as recombinational repair and translesion synthesis, have been widely studied[3,4,6–9]. However, little is known about how the replisome proceeds through protein–nucleic acid blocks. In particular, replication forks often collide with transcription complexes that translocate in the same (co-directional) or opposite (head-on) direction as the replisome[1,2,5]. In bacteria, the rate of replication ($\sim$600 nucleotides s^{-1}) is 12–30-fold greater than the rate of transcription (20–50 nucleotides s^{-1}) and there is no temporal separation between the two processes[1,2,10,11]. Thus, both head-on and co-directional collisions between the replisome and RNAP are probably frequent. Here we investigate the mechanism by which the *E. coli* replisome passes a RNAP that is co-directional with replication fork movement.

Essential genes and most transcription units in bacteria are encoded by the leading strand, which suggests a natural selection for co-directional collisions in the cell[1,5,12–14]. It therefore seems probable that cell survival requires the resolution of co-directional collisions in a manner that does not block fork progression. Indeed, *in vivo* studies in bacteria and eukaryotes indicate that co-directional transcription complexes do not impede replisome progression[1,2,15–20]. In contrast, head-on collisions predominately result in replication fork arrest and induce DNA recombination in bacteria and yeast[1,2,15–21]. In eukaryotes, replication fork barriers have evolved that prevent head-on collisions within highly expressed genes during S phase[1,2]. Furthermore, a recent study indicates that human cells also favour co-directional movement of the replisome with RNAP[22].

In a co-directional collision, the leading-strand DNA polymerase and RNAP use the same strand as a template (see Fig. 1a). The replicative helicase, DnaB, unwinds the DNA ahead of the *E. coli* replication fork by translocating on the opposite (lagging) strand. Thus, the helicase may continue past the RNAP in which case a physical interaction between the two co-directional polymerases on collision is almost certain. A question remains as to how the replication fork then bypasses a co-directional RNAP without collapsing. Previous *in vitro* studies of the bacteriophage T4 replisome indicate that a co-directional transcription complex poses no obstacle to the progression of the T4 replication fork[23–25]. These studies indicated that RNAP remains bound to the DNA during passage of the T4 replisome.

In this report we determine a new mechanism by which the *E. coli* replisome bypasses a co-directional transcription complex *in vitro*. We have used T7 RNAP as well as *E. coli* RNAP and found that the leading strand terminates after collision with RNAP, but in a notable transaction the replisome uses the messenger RNA as a primer to continue the leading strand. This process results in a discontinuity in the leading strand and therefore may explain why leading-strand synthesis is performed discontinuously *in vivo*[26,27].

Observation of co-directional collisions

The *E. coli* replicase—referred to as DNA polymerase III (Pol III) holoenzyme—is a multicomponent protein complex that performs rapid and highly processive replication of chromosomal DNA[28,29]. A co-directional RNAP may block the leading strand, and the current report focuses on leading-strand synthesis by omitting primase. The proteins that perform leading-strand synthesis are illustrated in Fig. 1a and include the following components: Pol III; the β-clamp, which confers processivity to Pol III; the clamp-loader, which assembles clamps at primed sites; and the DnaB helicase, which unwinds duplex DNA ahead of the replication fork.

We first investigated the effect of a co-directional bacteriophage T7 transcription complex on the progression of the *E. coli* replication fork. We constructed a 2.2-kb linear forked DNA template that supports replication from one end and includes a co-directional T7 RNAP promoter 1 kb downstream from the replication fork (Fig. 1b). T7 RNAP serves as a model system for multisubunit RNAPs such as *E. coli* RNAP and the basic mechanisms of transcription are identical between these enzymes[30].

During transcription initiation, RNAP binds to the promoter and unwinds DNA to form an open promoter complex. In Fig. 1c we addressed whether a co-directional T7 RNAP open promoter complex affects progression of the replication fork. The Pol III holoenzyme, DnaB and T7 RNAP were first pre-incubated with the 2.2-kb linear forked DNA in the presence of ATP, which results in the assembly of the replisome at the fork and a T7 RNAP open promoter

[1]The Rockefeller University, Howard Hughes Medical Institute, 1230 York Avenue, New York, New York 10021, USA.

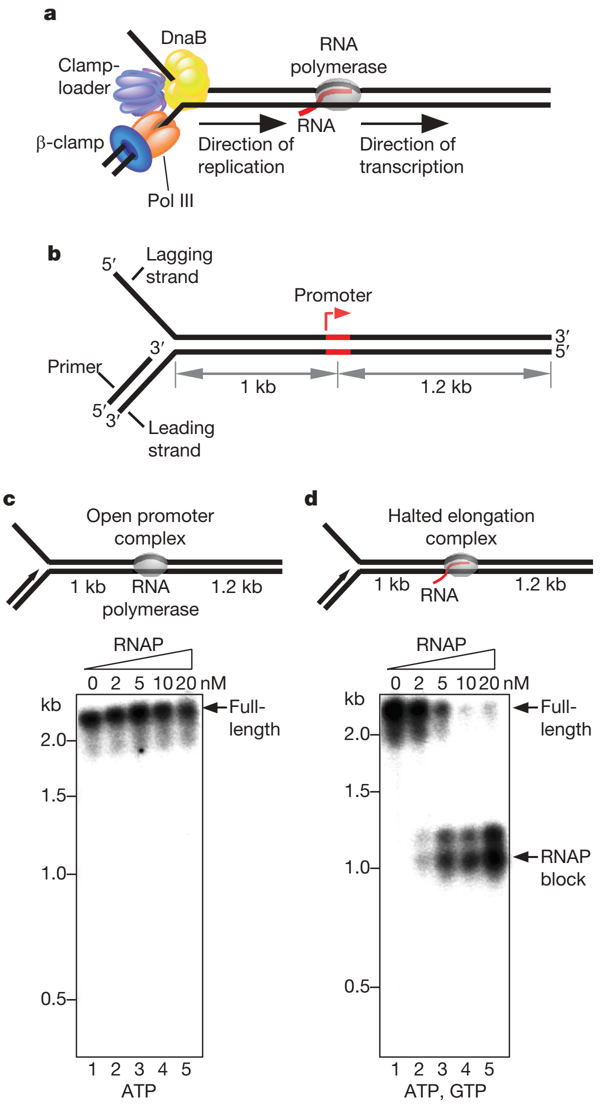


Figure 1 | Leading-strand synthesis is interrupted by a co-directional RNA polymerase. **a**, Schematic of replisome components and a co-directional RNAP. Replisome proteins include: Pol III core (orange), β-clamp (dark blue), DnaB (yellow) and the clamp-loader (light blue). Primase was omitted from reactions and the lagging-strand polymerase is not pictured. **b**, A 2.2-kb template was constructed that supports leading-strand synthesis and co-directional transcription. **c**, **d**, Leading-strand synthesis was performed in the presence of increasing concentrations of an RNAP open (**c**) and halted elongation (**d**) complex. Radiolabelled DNA products were analysed on alkaline agarose gels (**c**, **d**).

complex. Leading-strand synthesis was initiated by the addition of α-^{32}P-labelled deoxyribonucleoside triphosphates (dNTPs), and DNA products were analysed by electrophoresis in denaturing alkaline agarose gels. The results show that replisome progression is unaffected by the open promoter complex, as indicated by the appearance of only the full-length product (2.2 kb; Fig. 1c).

Once RNAP synthesizes a transcript ~10–12 nucleotides in length, it leaves the promoter and enters into a highly processive elongation complex[30–32]. Elongating RNAP often pauses or is arrested because of regulatory signals or lesions in the DNA[2,33]. Halted elongation complexes increase the probability of replisome–RNAP collisions in the cell, especially in strains that lack factors which revive or displace a halted RNAP[34]. To examine whether a halted co-directional T7 RNAP affects fork progression we added ATP and GTP, enabling RNAP to synthesize a 22-nucleotide transcript (Fig. 1d). If replisome advance is not blocked by a co-directional transcription complex, as indicated by *in vivo* studies, the full-length 2.2-kb product should still be observed. However, the result indicates that RNAP prevents the formation of a full-length leading-strand product and instead yields a 1-kb product—the distance to the halted RNAP (Fig. 1d). Notably we also observe a 1.2-kb product, which corresponds to the length of the DNA template downstream from the promoter. The formation of the 1-kb product suggests that leading-strand synthesis is terminated by the halted RNAP, but the 1.2-kb product suggests the unexpected possibility that the leading strand is reinitiated using the messenger RNA as a primer. This hypothesis predicts that the position of the RNAP along the template dictates the length of the two leading-strand products. Indeed, moving the promoter to a different position changes the lengths of the upstream and downstream leading-strand products accordingly (Supplementary Fig. 1).

Pol III uses an RNA transcript as a primer

To gain further evidence that Pol III uses the RNA transcript as a primer, we terminated the mRNA before initiating replication by adding 3′-deoxy-cytidine-ribonucleoside-triphosphate (3′dCTP), an RNA chain terminator that is incorporated by RNAP (Fig. 2a). The addition of 3′dCTP prevented synthesis of the downstream portion of the leading strand (1.2-kb DNA), but did not affect synthesis of the initial 1-kb product (Fig. 2a, compare lanes 1 and 2). Similar results were obtained using a template that includes the promoter at a different position (Supplementary Fig. 2). Next, we observed extension of the transcript by Pol III directly by labelling the RNA instead of the DNA (Fig. 2b). In this case α-^{32}P-GTP and α-^{32}P-ATP were added, which are incorporated into the 22-nucleotide transcript by RNAP before replication (Fig. 2b, lane 1). Initiating replication results in extension of the transcript to 1.2 kb, corresponding to the length of the DNA downstream from the halted RNAP (Fig. 2b, lane 2). These results confirm that the mRNA is extended by Pol III.

Next we used a 2.2-kb linear duplex without a forked junction to determine whether the replication proteins could assemble at the transcription bubble of a halted RNAP and extend the RNA to form a 1.2-kb product (Fig. 2c). However, no products were observed in the absence of a replication fork (Fig. 2c, lane 2). Therefore, collision of the replisome with the RNAP is required for Pol III extension of the transcript. The result in lane 2 (Fig. 2c) also demonstrates that RNAP is unable to form the 1.2-kb downstream product by misincorporating dNTPs.

Fate of the replisome and RNA polymerase

Because the replisome must collide with the transcription complex to gain access to the RNA, it is probable that the collision results in displacement of RNAP from the DNA. To test this we immobilized a His-tagged T7-RNAP-halted elongation complex to Ni^{2+} beads and addressed whether the DNA remains bound to the RNAP (pellet) or is released into solution (supernatant) after a co-directional collision (Fig. 3a). The transcription complex was first immobilized, and then unbound DNA and RNAP were removed by washing followed by the initiation of replication. Upstream (1 kb) and downstream (1.2 kb) products were only observed in the supernatant, indicating that RNAP is displaced by the replisome (Fig. 3a, left). Some full-length product was also observed, presumably owing to a fraction of transcription complexes that dissociated before replication. In the absence of replication, the DNA was analysed in a native agarose gel stained with ethidium bromide (Fig. 3a, right). In this case most of the DNA remained bound to RNAP (pellet), whereas only a small fraction of the DNA was released into the supernatant. These data support the conclusion that the replisome displaces a co-directional RNAP from the DNA.

Studies *in vivo* indicate that replication forks are not impeded by collisions with co-directional transcription complexes, suggesting that the replisome remains intact during bypass of a co-directional RNAP[1,2,15,16,20]. An important factor that determines the integrity of the replication fork is whether the replicative helicase, DnaB, remains associated with the lagging strand (see Fig. 1a). To determine whether DnaB dissociates from the replisome during bypass of a co-directional RNAP we assembled the replisome and a halted T7

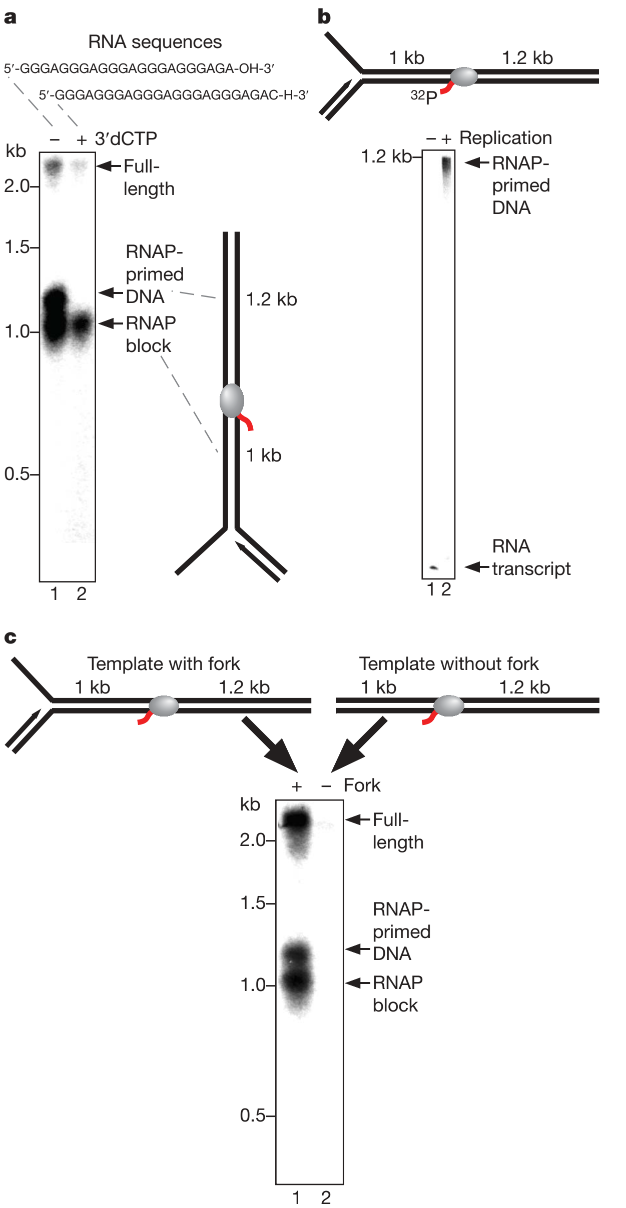


Figure 2 | The replisome extends the transcript of a co-directional RNA polymerase. a, A co-directional collision of the replisome with a halted RNAP was performed. Extension of the RNA was permitted (lane 1) or blocked (lane 2) by the addition of RNA chain terminator 3′dCTP. RNA sequences are indicated. **b,** The transcript was radiolabelled by the addition of α-^{32}P-GTP and α-^{32}P-ATP and analysed by urea-PAGE before (lane 1) and after (lane 2) replication. **c,** A co-directional collision was performed on a template that either includes (lane 1) or lacks (lane 2) a fork structure. Radiolabelled DNA products were analysed on alkaline agarose gels (**a**, **c**).

RNAP on a biotinylated template in the presence of ATP and GTP, and then immobilized the DNA to streptavidin beads (Fig. 3b). Excess unbound DnaB and Pol III holoenzyme were removed by washing. Replication was then initiated after the addition of dNTPs, the β-clamp and single-strand binding protein (SSB), and radiolabelled DNA products were analysed on an alkaline agarose gel. The results show that both 1- and 1.2-kb products were formed, indicating that the replisome can bypass a co-directional RNAP without dissociating from DNA (Fig. 3b, lane 2). In a control reaction RNAP was omitted, which resulted in only full-length product (Fig. 3b, lane 1). A further control reaction demonstrates that replication proteins do not adhere to the beads after washing (Supplementary Fig. 3). To ensure that DnaB is a necessary participant in these reactions, the experiment was repeated but the helicase

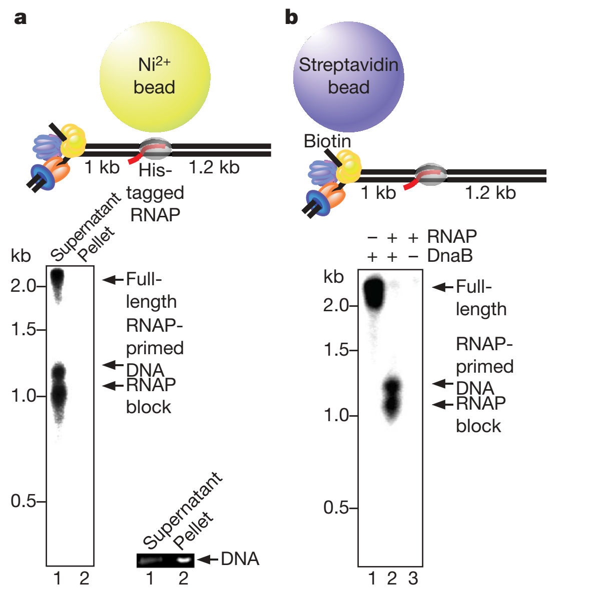


Figure 3 | The replisome remains intact and displaces a co-directional RNA polymerase from the DNA. a, A His-tagged RNAP halted elongation complex was assembled and immobilized to Ni^{2+} beads. Excess RNAP and DNA were removed by washing followed by replication initiation. Supernatant and pellet fractions were analysed on an alkaline agarose gel (left). A His-tagged RNAP-halted elongation complex was assembled and immobilized as in the left panel; however, replication was not initiated. Supernatant and pellet fractions were analysed on a non-denaturing agarose gel stained with ethidium bromide (bottom right). **b,** Leading-strand synthesis was performed in solid-phase after the removal of excess Pol III* and DnaB (lanes 1 and 2) in either the presence (lanes 2 and 3) or the absence (lane 1) of a co-directional halted RNAP. DnaB was omitted in lane 3. Radiolabelled DNA products were analysed on an alkaline agarose gel.

was omitted (Fig. 3b, lane 3). The absence of products in lane 3 indicates that DnaB is required for leading-strand synthesis, as expected. These results indicate that the only proteins required for replisome bypass of a co-directional RNAP are those that are present at the replication fork, and that the replisome bypasses RNAP without collapsing.

Replisome bypass of *E. coli* RNA polymerase

Although T7 RNAP serves as an important model enzyme, the multi-subunit *E. coli* RNAP could conceivably behave differently. Therefore, we examined the replisome for the ability to bypass a halted *E. coli* RNAP (Fig. 4). We constructed a linear 3.5-kb DNA that includes the strong *E.coli* RNAP T7A1 promoter 1.1 kb downstream from the replication fork and a biotin at the downstream edge. A halted *E. coli* RNAP elongation complex was first assembled by the addition of *E. coli* RNAP σ^{70} holoenzyme, ApU, GTP, CTP and ATP, which limits RNA synthesis to 20 nucleotides. The DNA was then immobilized to streptavidin beads and washed with high salt to remove non-specific RNAP–DNA complexes. The fork was ligated to the DNA followed by initiation of replication. Similar to experiments using a halted T7 RNAP, we observe replication products corresponding to the lengths of the template upstream (1.1 kb) and downstream (2.4 kb) from the promoter as well as some full-length product (Fig. 4, lane 2). The percentage of full-length product (33%) corresponds relatively well to the number of promoters unoccupied by RNAP (24%; Supplementary Fig. 4). Omitting RNAP from the reaction resulted in only full-length product (Fig. 4, lane 1). Finally, we observe Pol III extension of the 20-nucleotide *E. coli* RNAP transcript directly by labelling the RNA (Supplementary Fig. 5). These data indicate that the replisome can bypass a halted co-directional *E. coli* RNAP by using the transcript as a primer to continue the leading strand as observed using the T7 RNAP.

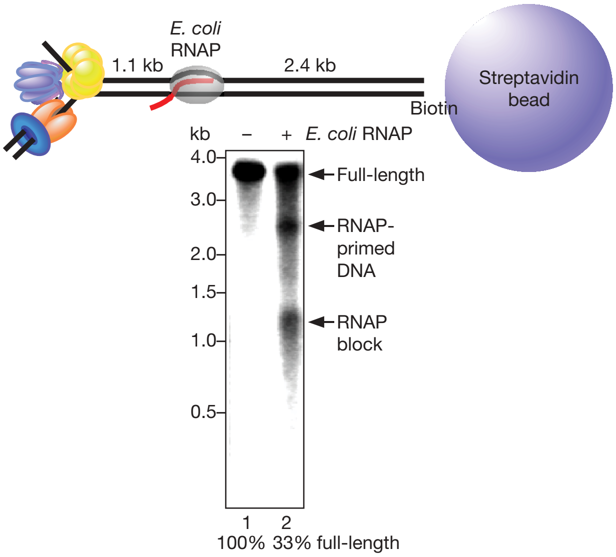


Figure 4 | Replisome bypass of a co-directional *E. coli* RNAP elongation complex. Leading-strand synthesis was performed in solid phase on a 3.5-kb template that either includes (lane 2) or lacks (lane 1) a co-directional halted *E. coli* RNAP elongation complex. Radiolabelled DNA products were analysed on an alkaline agarose gel. The percentage of full-length product is indicated and was calculated as described in the Methods.

Discussion

Here we demonstrate that leading-strand synthesis is terminated after colliding with a co-directional RNAP, but can then be reinitiated by using the mRNA as a primer. A model of this mechanism is presented in Fig. 5. We propose that RNAP is displaced from the DNA by the leading-strand polymerase, whereas DnaB remains bound to the lagging strand. The leading-strand polymerase hops over the mRNA by remaining bound to the clamp-loader which assembles a new clamp at the 3′ terminus of the RNA–DNA hybrid. Pol III then binds to the newly assembled clamp and extends the transcript, leaving behind a nick or gap in the leading strand. The RNA can then be excised and replaced by DNA in a similar repair reaction as occurs during the maturation of Okazaki fragments.

The scheme hypothesized in Fig. 5 has precedent in synthesis of the lagging strand in which Pol III rapidly hops from a clamp on a completed Okazaki fragment to a newly assembled clamp on a RNA–DNA hybrid every few seconds. Collision of the lagging strand polymerase with the 5′ terminus of an Okazaki fragment triggers the release of Pol III from the clamp[35]. Thus, the hopping of the leading-strand polymerase proposed in Fig. 5 may be initiated by a similar collision mechanism. During lagging-strand synthesis RNA primers are made by primase. Here the RNA primers are provided by RNAP on the leading strand. Primase activity on the leading strand is probably low because it requires stimulation by DnaB on the lagging strand.

In vivo, replication forks presumably encounter co-directional RNAPs that have synthesized long transcripts. We are at present investigating the consequence of replisome collision with co-directional transcription complexes farther downstream from the promoter. Replisome takeover of long transcripts in the cell might trigger translational regulatory mechanisms such as the trans-translation system, which removes stalled ribosomes from truncated mRNA and targets the mRNA for degradation[36].

Synthesis of the leading strand is predominately viewed as a continuous process. This view is mostly on the basis of *in vitro* studies that lack impediments to the replication fork. In contrast, several *in vivo* studies demonstrate that the leading strand is synthesized in a discontinuous fashion even as far back as Okazaki's original work[26,27,37–45]. One source of leading-strand interruptions could be due to replication fork collapse, because restart mechanisms that reactivate the fork involve new primers and thus produce single-strand gaps[4,46,47]. Here we provide a new explanation for leading-strand interruptions in which a replication fork simply recruits the 3′ terminus of the mRNA to continue leading-strand synthesis after a collision with RNAP. These protein dynamics emphasize the remarkable plasticity of the moving replisome apparatus, and underscore a driving force during evolution that has enabled replication machines to efficiently deal with obstacles along the path of chromosome duplication.

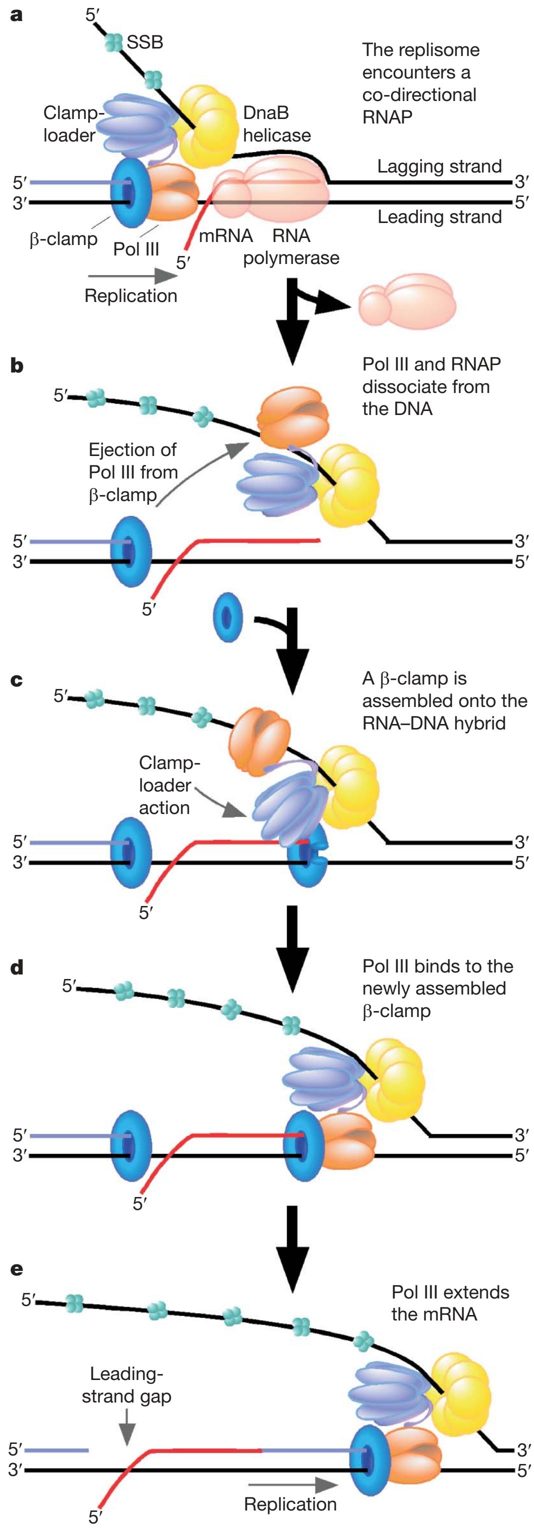


Figure 5 | Model of replisome bypass of a co-directional RNA polymerase. **a**, The replisome encounters a co-directional RNAP. **b**, RNAP is displaced from the DNA. The lagging-strand polymerase dissociates from the β-clamp and DNA but remains bound to the clamp-loader. DnaB remains bound to the lagging strand. **c**, The clamp-loader assembles a new β-clamp at the 3′ terminus of the RNA–DNA hybrid. **d**, The leading-strand polymerase binds to the newly assembled β-clamp. **e**, The leading-strand polymerase extends the mRNA leaving behind a nick or gap in the leading strand.

METHODS SUMMARY

DNA templates. Linear forked DNA was prepared in a similar fashion to a previous study[48]. For 2.2-kb DNA, pPK7 (ref. 31) was digested with BsaI followed by ligation in the presence of excess complementary forked DNA that was pre-annealed by mixing oligonucleotides RP25, RP26 and RP33 together, followed by boiling and slow cooling to room temperature. The 2.2-kb DNA without a fork was prepared by digesting pPK7 with BsaI. For the 10.5-kb DNA, pRSF2 was digested with SapI and then ligated in the presence of excess forked DNA (RP25, RP26 and RP10). For the 3.5-kb DNA, PCR was performed using pRP50 as a template and the primers RP64B and RP65. The PCR product was purified, digested with SapI and then ligated in the presence of excess forked DNA (RP25, RP26 and RP10). The ligation products were purified by gel filtration followed by phenol extraction and ethanol precipitation.

Leading-strand synthesis. DnaB (44.8 pmol, as hexamer) was incubated with 1.5 nM final concentration of linear forked DNA in 15 μl of buffer A (20 mM Tris-HCl, pH 7.5), 8 mM $MgCl_2$, 0.5 mM EDTA, 5 mM dithiothreitol, 10% glycerol) for 15 s at 23 °C. Then 488 fmol of Pol III* (Pol III holoenzyme minus β-clamp), 1.5 pmol of β-clamp, 2 mM ATP and 60 μM each of dGTP and dATP were added to a volume of 20 μl and incubated a further 5 min at 23 °C. Replication was initiated after adding 1 μg SSB and α-^{32}P-dTTP and α-^{32}P-dCTP (specific activity, 3,000–5,000 c.p.m. pmol^{-1}) to a final volume of 25 μl. Reactions were terminated after 10 min on adding 5 μl of 120 mM EDTA and 3% SDS. All experiments (except where indicated) used Pol III* reconstituted from pure subunits and an ε-mutant that abolishes 3′–5′ exonuclease activity[49]. Radiolabelled products were analysed in alkaline agarose gels.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 August; accepted 6 October 2008.
Published online 19 November 2008.

1. Mirkin, E. V. & Mirkin, S. M. Replication fork stalling at natural impediments. *Microbiol. Mol. Biol. Rev.* **71,** 13–35 (2007).
2. Rudolph, C. J., Dhillon, P., Moore, T. & Lloyd, R. G. Avoiding and resolving conflicts between DNA replication and transcription. *DNA Repair (Amst.)* **6,** 981–993 (2007).
3. Cox, M. M. *et al.* The importance of repairing stalled replication forks. *Nature* **404,** 37–41 (2000).
4. Cox, M. M. Recombinational DNA repair of damaged replication forks in *Escherichia coli*: questions. *Annu. Rev. Genet.* **35,** 53–82 (2001).
5. Brewer, B. J. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53,** 679–686 (1988).
6. Jarosz, D. F., Beuning, P. J., Cohen, S. E. & Walker, G. C. Y-family DNA polymerases in *Escherichia coli. Trends Microbiol.* **15,** 70–77 (2007).
7. Goodman, M. F. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu. Rev. Biochem.* **71,** 17–50 (2002).
8. Tippin, B., Pham, P. & Goodman, M. F. Error-prone replication for better or worse. *Trends Microbiol.* **12,** 288–295 (2004).
9. Lusetti, S. L. & Cox, M. M. The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. *Annu. Rev. Biochem.* **71,** 71–100 (2002).
10. Kornberg, A. & Baker, T. A. *DNA Replication* 2nd edn 246 (W. H. Freeman & Co, 1992).
11. Breier, A. M., Weier, H. U. & Cozzarelli, N. R. Independence of replisomes in *Escherichia coli* chromosomal replication. *Proc. Natl Acad. Sci. USA* **102,** 3942–3947 (2005).
12. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).
13. Rocha, E. P. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genet.* **34,** 377–378 (2003).
14. Rocha, E. P. & Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31,** 6570–6577 (2003).
15. Mirkin, E. V. & Mirkin, S. M. Mechanisms of transcription-replication collisions in bacteria. *Mol. Cell. Biol.* **25,** 888–895 (2005).
16. Wang, J. D., Berkmen, M. B. & Grossman, A. D. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in *Bacillus subtilis. Proc. Natl Acad. Sci. USA* **104,** 5608–5613 (2007).
17. Prado, F. & Aguilera, A. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *EMBO J.* **24,** 1267–1276 (2005).
18. Kobayashi, T. The replication fork barrier site forms a unique structure with Fob1p and inhibits the replication fork. *Mol. Cell. Biol.* **23,** 9178–9188 (2003).
19. Deshpande, A. M. & Newlon, C. S. DNA replication fork pause sites dependent on transcription. *Science* **272,** 1030–1033 (1996).
20. French, S. Consequences of replication fork movement through transcription units *in vivo. Science* **258,** 1362–1365 (1992).
21. Vilette, D., Ehrlich, S. D. & Michel, B. Transcription-induced deletions in plasmid vectors: M13 DNA replication as a source of instability. *Mol. Gen. Genet.* **252,** 398–403 (1996).
22. Huvet, M. *et al.* Human gene organization driven by the coordination of replication and transcription. *Genome Res.* **17,** 1278–1285 (2007).
23. Liu, B. & Alberts, B. M. Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science* **267,** 1131–1137 (1995).
24. Liu, B. *et al.* The DNA replication fork can pass RNA polymerase without displacing the nascent transcript. *Nature* **366,** 33–39 (1993).
25. Liu, B., Wong, M. L. & Alberts, B. A transcribing RNA polymerase molecule survives DNA replication without aborting its growing RNA chain. *Proc. Natl Acad. Sci. USA* **91,** 10660–10664 (1994).
26. Ogawa, T. & Okazaki, T. Discontinuous DNA replication. *Annu. Rev. Biochem.* **49,** 421–457 (1980).
27. Wang, T. C. Discontinuous or semi-discontinuous DNA replication in *Escherichia coli*? *Bioessays* **27,** 633–636 (2005).
28. Johnson, A. & O'Donnell, M. Cellular DNA replicases: components and dynamics at the replication fork. *Annu. Rev. Biochem.* **74,** 283–315 (2005).
29. Pomerantz, R. T. & O'Donnell, M. Replisome mechanics: insights into a twin DNA polymerase machine. *Trends Microbiol.* **15,** 156–164 (2007).
30. Steitz, T. A. The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. *Curr. Opin. Struct. Biol.* **14,** 4–9 (2004).
31. Mentesana, P. E., Chin-Bow, S. T., Sousa, R. & McAllister, W. T. Characterization of halted T7 RNA polymerase elongation complexes reveals multiple factors that contribute to stability. *J. Mol. Biol.* **302,** 1049–1062 (2000).
32. Jiang, M., Rong, M., Martin, C. & McAllister, W. T. Interrupting the template strand of the T7 promoter facilitates translocation of the DNA during initiation, reducing transcript slippage and the release of abortive products. *J. Mol. Biol.* **310,** 509–522 (2001).
33. Uptain, S. M., Kane, C. M. & Chamberlin, M. J. Basic mechanisms of transcript elongation and its regulation. *Annu. Rev. Biochem.* **66,** 117–172 (1997).
34. Trautinger, B. W., Jaktaji, R. P., Rusakova, E. & Lloyd, R. G. RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Mol. Cell* **19,** 247–258 (2005).
35. Stukenberg, P. T., Turner, J. & O'Donnell, M. An explanation for lagging strand replication: polymerase hopping among DNA sliding clamps. *Cell* **78,** 877–887 (1994).
36. Keiler, K. C. Biology of *trans*-translation. *Annu. Rev. Microbiol.* **62,** 133–151 (2008).
37. Okazaki, R. *et al.* Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl Acad. Sci. USA* **59,** 598–605 (1968).
38. Sternglanz, R., Wang, H. F. & Donegan, J. J. Evidence that both growing DNA chains at a replication fork are synthesized discontinuously. *Biochemistry* **15,** 1838–1843 (1976).
39. Pauling, C. & Hamm, L. Properties of a temperature-sensitive, radiation-sensitive mutant of *Escherichia coli*. II. DNA replication. *Proc. Natl Acad. Sci. USA* **64,** 1195–1202 (1969).
40. Gottesman, M. M., Hicks, M. L. & Gellert, M. Genetics and function of DNA ligase in *Escherichia coli. J. Mol. Biol.* **77,** 531–547 (1973).
41. Konrad, E. B., Modrich, P. & Lehman, I. R. Genetic and enzymatic characterization of a conditional lethal mutant of *Escherichia coli* K12 with a temperature-sensitive DNA ligase. *J. Mol. Biol.* **77,** 519–529 (1973).
42. Okazaki, R., Arisawa, M. & Sugino, A. Slow joining of newly replicated DNA chains in DNA polymerase I-deficient *Escherichia coli* mutants. *Proc. Natl Acad. Sci. USA* **68,** 2954–2957 (1971).
43. Olivera, R. M. & Bonhoeffer, E. Replication of *Escherichia coli* requires DNA polymerase I. *Nature* **250,** 513–514 (1974).
44. Wang, T. C. & Smith, K. C. Discontinuous DNA replication in a *lig-7* strain of *Escherichia coli* is not the result of mismatch repair, nucleotide-excision repair, or the base-excision repair of DNA uracil. *Biochem. Biophys. Res. Commun.* **165,** 685–688 (1989).
45. Wang, T. C. & Chen, S. H. Okazaki DNA fragments contain equal amounts of lagging-strand and leading-strand sequences. *Biochem. Biophys. Res. Commun.* **198,** 844–849 (1994).
46. Heller, R. C. & Marians, K. J. Replication fork reactivation downstream of a blocked nascent leading strand. *Nature* **439,** 557–562 (2006).
47. Heller, R. C. & Marians, K. J. Replisome assembly and the direct restart of stalled replication forks. *Nature Rev. Mol. Cell Biol.* **7,** 932–943 (2006).
48. Heller, R. C. & Marians, K. J. The disposition of nascent strands at stalled replication forks dictates the pathway of replisome loading during restart. *Mol. Cell* **17,** 733–743 (2005).
49. McInerney, P. & O'Donnell, M. Functional uncoupling of twin polymerases: mechanism of polymerase dissociation from a lagging-strand block. *J. Biol. Chem.* **279,** 21543–21551 (2004).
50. Severinov, K., Mooney, R., Darst, S. A. & Landick, R. Tethering of the large subunits of *Escherichia coli* RNA polymerase. *J. Biol. Chem.* **272,** 24137–24140 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to W. T. McAllister and R. Castagna for providing T7 RNAP, and to S. Darst and L. Westblade for providing *E. coli* RNAP proteins and plasmids. This work was supported by a grant from the National Institutes of Health (M.O.D.) and by a Marie-Josee and Henry Kravis Fellowship at the Rockefeller University (R.T.P.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.O.D. (odonnell@mail.rockefeller.edu).

METHODS

Co-directional collision of the replisome with T7 RNAP. Leading-strand synthesis was performed as described in Methods Summary except for the following additions: 20 nM (or as specified) T7 RNAP was added along with Pol III* and β-clamp. Then, either 3 μM GTP or a mixture of 3 μM GTP and 1 μM UTP was added along with DnaB to assemble a T7 RNAP-halted elongation complex on the 2.2-kb and 10.5-kb templates, respectively. Assembly of the T7 RNAP open promoter complex required no additional NTPs. Twenty micromolar 3′dCTP (TriLink) was added along with DnaB in the experiments shown in Fig. 2a and Supplementary Fig. 2. Supplementary Figs 1 and 2 included a 10.5-kb forked DNA template instead of the 2.2-kb DNA substrate and a mixture of 3-μM GTP and 1-μM UTP was added along with DnaB.

In the experiment of Fig. 2b, 20 reactions were pooled, dNTPs were unlabelled, and α-^{32}P-GTP and α-^{32}P-ATP were added along with DnaB and Pol III*, respectively. Reactions were terminated by removing nucleotides through centrifugation over G-25 spin columns (Roche) followed by phenol extraction and ethanol precipitation with 10 μg of carrier DNA. Precipitated nucleic acid was resuspended in 10 μl of 20 mM Tris-HCl, pH 8.5, and then mixed with 10 μl of 90% (w/v) formamide and 50 mM EDTA. Samples were boiled and analysed on an 8% urea-polyacrylamide gel.

Collision of the replisome with T7 RNAP immobilized to beads. A final concentration of 130 nM of His-tagged T7 RNAP was incubated with a final concentration of 10 nM of 2.2-kb linear forked DNA along with 300 μM GTP and 100 μM ATP in 25 μl of buffer A for 5 min at room temperature. Thirty microlitres of Ni^{2+} magnetic coated beads (Promega) was added for a further 5 min. Next, the beads were washed three times with 100 μl of buffer A. Leading-strand synthesis was then performed as described in Methods Summary except for the following modifications: 1 and 3 pmol of Pol III* and β-clamp were added, respectively. After the reaction was terminated, the supernatant (25 μl total volume) was removed for analysis. The beads were then washed twice with 100 μl of buffer A. The pellet fraction was removed from the beads by the addition of 0.5 M imidazole and 100 mM EDTA in a total volume of 25 μl for 5 min at room temperature. Equal volumes of supernatant and pellet fractions were analysed on an alkaline agarose gel. In the absence of leading-strand synthesis (Fig. 3a, right), the supernatant and pellet fractions were analysed on a native agarose gel stained with ethidium bromide.

Collision of single replisome particles with T7 RNAP on immobilized DNA. Where indicated, 44.8 pmol DnaB was incubated with 5 nM final concentration of 2.2-kb linear forked DNA, which was biotinylated at the 5′ terminus of the lagging strand, in 15 μl of buffer A for 15 s at 23 °C. T7 RNAP (20 μM) and GTP (3 μM) were added (where indicated) along with DnaB. Pol III* (841 fmol; including wild-type ε) and β-clamp (5 pmol) were then added along with 2 mM ATP and 60 μM each of dGTP and dATP to a volume of 20 μl for a further 5 min. Reactions were mixed with 20 μl of streptavidin-coated magnetic beads (Invitrogen) pre-washed with buffer A for 10 min at 23 °C. Beads were washed three times with 100 μl of buffer A along with 60 μM each of dGTP and dATP, 2 mM ATP, 5 pmol of β-clamp and, where indicated, 20 nM T7 RNAP along with 3 μM GTP. Beads were resuspended in 20 μl of their respective wash buffers (with or without T7 RNAP and GTP) and replication was initiated as described in Methods Summary. Reactions were terminated after 20 min by the addition of 5 μl of 120 mM EDTA and 3% SDS. Beads were boiled and the supernatant was removed for gel analysis. Beads were then treated with proteinase K in 10 μl of 10 mM Tris-HCl, pH 7.5, 5 mM EDTA, 1% SDS for 30 min at 50 °C to remove residual DNA from the solid support. The supernatant was pooled and radiolabelled DNA was analysed on a 1.2% alkaline agarose gel. The experiment in Supplementary Fig. 2 was performed in a similar fashion, except that T7 RNAP and GTP were omitted and biotinylated DNA was either pre-incubated with DnaB or added along with SSB and dNTPs as indicated.

Co-directional collision of the replisome with an *E. coli* RNAP elongation complex. A final concentration of 500 nM of *E. coli* RNAP σ^{70} holoenzyme was mixed with a final concentration of 5 nM of a 3.5-kb DNA in 100 μl of buffer A for 10 min at 37 °C. One-hundred micromolar ApU and 40 μM each of GTP and ATP were added for a further 10 min at 37 °C. Two-hundred microlitres of streptavidin magnetic coated beads (Invitrogen) was added for a further 10 min at room temperature. The beads were washed five times with 0.9 ml of buffer A containing 0.75 M NaCl, 200 μg ml^{-1} heparin, and 20 μg ml^{-1} single-stranded DNA. Next, the beads were washed twice with 0.9 ml of buffer A and then resuspended in 100 μl of New England Biolabs buffer 4, and 10 units of Sap I (New England Biolabs) was added for 10 min at 37 °C. The beads were washed three times with 0.9 ml of buffer A and then resuspended in 50 μl of Quick Ligation reaction buffer (New England Biolabs). Two microlitres of Quick T4 ligase (New England Biolabs) was added along with 6 nM final concentration of pre-annealed forked DNA (RP10, RP22 and RP25) for 10 min at room temperature. The beads were washed three times with 0.9 ml of buffer A. Next, leading-strand synthesis was performed as described in Methods Summary except ten reactions were pooled. The beads were boiled after the reaction was terminated and the supernatant was purified using the Qiagen PCR Cleanup kit. Purified radiolabelled DNA products were analysed on an alkaline agarose gel. The percentage of full-length product was calculated using the equation: $I_{FL}[I_{FL}+(I_B 3.18)]^{-1}\times 100$, in which I_{FL} denotes the intensity of full-length product, and I_B represents the intensity of the replication block. The factor 3.18 corrects for the amount of full-length product that would have been formed relative to the intensity of the replication block (I_B) and was calculated by dividing the length of the full-length product (3.5 kb) by the length of the blocked product (1.1 kb). The occupancy of promoters bound by *E. coli* RNAP in Supplementary Fig. 4 was determined by XhoI restriction digest of the immobilized 3.5 kb DNA in the absence of leading-strand synthesis either with or without the addition of *E. coli* RNAP.

Pol III extension of a co-directional *E. coli* RNAP transcript. Leading-strand synthesis was performed as described in Methods Summary except for the following modifications: 30 reactions were pooled and performed at 37 °C. Forty micromolar each of ApU, GTP and CTP were added along with DnaB, which was incubated with DNA for 30 s rather than 15 s. A final concentration of 50 nM of *E. coli* RNAP σ^{70} holoenzyme was added 2 min after the addition of Pol III* and β-clamp. α-^{32}P-dNTPs were omitted and α-^{32}P-GTP and α-^{32}P-ATP were added along with DnaB and Pol III*, respectively. Reactions were terminated by removing nucleotides through centrifugation over G-25 spin columns (Roche) followed by phenol extraction and ethanol precipitation with 10 μg of carrier DNA and 30 μg of glycogen. Precipitated nucleic acid was resuspended in 5 μl of 20 mM Tris-HCl, pH 8.5, and then mixed with 5 μl of 90% (w/v) formamide and 50 mM EDTA. Samples were boiled and analysed on an 8% urea-polyacrylamide gel.

Proteins. Replication proteins were expressed, purified and reconstituted as previously described[49]. T7 RNAP and *E. coli* RNAP core were gifts from W. T. McAllister and S. Darst, respectively. σ^{70} was expressed from pET21aEc(His)6PPXσ^{70} which was a gift from S. Darst. His-tagged σ^{70} was purified on a Ni^{2+} column and then concentrated on a Mono-Q column.

DNA. pPK7 (ref. 31) was a gift from W. T. McAllister. pRP50 was derived from pRL706 (ref. 50) which includes the *rpoB* gene of *E. coli*. The T7A1 promoter sequence was inserted into the *rpoB* gene by ligation of pre-annealed oligonucleotides RP35 and RP36 to ClaI-digested pRL706 to form pRP50. pRSF2 was constructed by inserting a 6.6-kb synthetic gene into a pRSFDuet-1 vector digested with NdeI and BglII. Oligonucleotide sequences were RP10, 5′-phosphate-AGCTGAGACCGCAATACGGATAAGGGCTGAGCACGTCCTGCGATCTGCAGCCTGCCAGAATCTGTG-3′; RP25, 5′-OH-CACAGATTCTGGCAGGCTGCAGATCGC-3′; RP22, 5′-phosphatase-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAGCCCTTATCCGTATTGCGGTCTCA-3′; RP26, 5′-biotin-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAGCCCTTATCCGTATTGCGGTCTCA-3′; RP33, 5′-phosphate-CGGTTGAGACCGCAATACGGATAAGGGCTGAGCACGTCCTGCGATCTGCAGCCTGCCAGAATCTGTG-3′; RP35, 5′-OH-CGGACGTTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGAGGGACACGGCGAATTCTCGAG-3′; RP36, 5′-OH-CGCTCGAGAATTCGCCGTGTCCCTCTCGATGGCTGTAAGTATCCTATAGGTTAGACTTTAAGTCAACGTC-3′; RP64B, 5′-biotin-AACCGGTGGAACGCGCGTGC; RP65, 5′-OH-TTTCATCTGCTCTTCCGCTTCCACCGCCTTGGCGAACCGGTG-3′.

Equipment and settings. All gels with the exception of that in Supplementary Fig. 2, were analysed with a phosphorimager using a 200 pixel per inch resolution setting. Gel images were then converted to tiff format and adjusted for contrast using Adobe Photoshop software version 9. Image sections were then selected, copied and pasted into a Canvas version 9 file. Pasted selections were converted into images and cropped further using Canvas. The gel in Supplementary Fig. 2 was photographed while exposed to ultraviolet light. The digital image was cropped and adjusted for contrast using Adobe Photoshop version 9. The image was then selected, copied and pasted into a Canvas file. All other image art was produced using Canvas with the exception of Supplementary Fig. 2, which includes a digital graph that was created using Excel.

50. Severinov, K., Mooney, R., Darst, S. A. & Landick, R. Tethering of the large subunits of *Escherichia coli* RNA polymerase. *J. Biol. Chem.* **272,** 24137–24140 (1997).

Vol 456 | 11 December 2008 | doi:10.1038/nature07574

Strong water absorption in the dayside emission spectrum of the planet HD 189733b

Carl J. Grillmair[1], Adam Burrows[2], David Charbonneau[3], Lee Armus[1], John Stauffer[1], Victoria Meadows[4], Jeffrey van Cleve[5], Kaspar von Braun[6] & Deborah Levine[1]

Recent observations of the extrasolar planet HD 189733b did not reveal the presence of water in the emission spectrum of the planet[1]. Yet models of such 'hot-Jupiter' planets predict an abundance of atmospheric water vapour[2]. Validating and constraining these models is crucial to understanding the physics and chemistry of planetary atmospheres in extreme environments. Indications of the presence of water in the atmosphere of HD 189733b have recently been found in transmission spectra[3,4], where the planet's atmosphere selectively absorbs the light of the parent star, and in broadband photometry[5]. Here we report the detection of strong water absorption in a high-signal-to-noise, mid-infrared emission spectrum of the planet itself. We find both a strong downturn in the flux ratio below 10 μm and discrete spectral features that are characteristic of strong absorption by water vapour. The differences between these and previous observations are significant and admit the possibility that predicted planetary-scale dynamical weather structures[6] may alter the emission spectrum over time. Models that match the observed spectrum and the broadband photometry suggest that heat redistribution from the dayside to the nightside is weak. Reconciling this with the high nightside temperature[7] will require a better understanding of atmospheric circulation or possible additional energy sources.

The extrasolar giant planet HD 189733b (ref. 8) is the most easily observable of the known transiting[9] extrasolar planets and has recently been the subject of intense scrutiny by both ground-based and space-based observatories. Owing to its relative proximity (19 pc), its high temperature (1,201 K (ref. 10)) and its large size relative to its parent star (ratio of radii, 0.155 (ref. 10)), the light from the planet can easily be distinguished in the infrared as a change in flux when the planet is observed during and out of secondary eclipse (when the planet is hidden from view by the parent star). This favourable orbital alignment has enabled highly accurate measurements of temperature variations with planetary longitude[7] as well as a preliminary infrared emission spectrum of the dayside atmosphere[1].

Broadband filter observations of HD 189733b during primary transit using the Infrared Array Camera (IRAC) on NASA's Spitzer Space Telescope recently yielded a flux deficit at 3.6 μm that was interpreted as absorption by water vapour[4]. However, an independent analysis of the same data concluded that the uncertainties in the observations are too large to support that interpretation[11]. Spectrophotometric observations of HD 189733b during primary transit with the Near Infrared Camera and Multi-Object Spectrometer on NASA's Hubble Space Telescope showed indications of both water vapour and methane[3], whereas observations at visible wavelengths made using Hubble's Advanced Camera for Surveys indicate a smooth but featureless trend towards smaller radii at longer wavelengths[12], suggesting the presence of small-particle hazes. Detailed modelling[13] of mid-infrared broadband measurements made during secondary eclipse[5] suggest that both water and carbon monoxide are present in the planet's emission spectrum.

We observed HD 189733b with the Spitzer Space Telescope's Infrared Spectrograph[14] (IRS) during ten secondary eclipses between mid-June and mid-December 2007. Each exposure series was begun three hours before mid-eclipse and ended three hours after mid-eclipse. These data extend from 5 to 14 μm and provide the first mid-infrared spectrum below 7.5 μm. This is important because the strongest spectral signatures of absorption by water vapour and other molecular species occur between 5 and 10 μm. Removal of the effects of drifts, detector latencies and telescope pointing oscillations has been described elsewhere[1]. In essence, our analysis measures the planetary light by subtracting the spectra obtained during secondary eclipse from the spectra taken before and after eclipse. In practice, we fit light curves to the time series data in each wavelength bin to make use of all available data, including the ingress and egress portions of the eclipse. The uncertainties are dominated by the short duration of the planet's passage behind the star (~1.5 h for HD 189733b), during which we can measure the spectrum of the parent star in isolation. The Spitzer Space Telescope is currently the only facility capable of carrying out these observations.

In Fig. 1 we show the planet/star flux ratio as a function of wavelength. Also shown are the 3.6-, 4.5-, 5.8-, 8.0-, 16.0- and 24.0-μm broadband flux ratio measurements made using all three of the Spitzer Space Telescope's detectors (IRAC, IRS and the Multiband Imaging Photometer for Spitzer)[5,15] as well as an upper limit[16] on the flux ratio at 2.2 μm obtained with the Near-Infrared Spectrograph on the Keck II telescope (W. M. Keck Observatory). The shape of the spectrum agrees quite well with that of the broadband measurements, although the 8.0-μm point lies slightly above the spectrum. The light-curve analyses applied to the two data sets are very similar (in as much as we treat the spectrograph as 128 independent photometers), and we know of no detector peculiarities that could give rise to a systematic offset. On the other hand, modestly significant differences have been noted in separate IRAC observations of the planet's 8.0-μm emission[5], suggesting that the planet's global dayside emission may actually be changing with time.

Also shown in Fig. 1 are three detailed model atmospheres, all incorporating equilibrium molecular abundances at solar elemental metallicity. The model spectra have been smoothed and rely on three different assumed values of the heat redistribution parameter, $P_n = 0.1$, 0.15, 0.3, and two values of the opacity, $\kappa_e = 0.0$, 0.035 $cm^2\ g^{-1}$, of a possible upper-atmosphere absorber that could be responsible for creating a slightly hotter upper atmosphere owing to anomalous absorption of the incident stellar light[17]. The 3.6-μm/

[1]Spitzer Science Center, 1200 East California Boulevard, Pasadena, California 91125, USA. [2]Department of Astrophysical Sciences, Princeton University, Peyton Hall, Ivy Lane, Princeton, New Jersey 08544, USA. [3]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [4]Department of Astronomy, University of Washington, Box 351580, Seattle, Washington 98195, USA. [5]Ball Aerospace & Technologies Corporation, PO Box 1062, Boulder, Colorado 80306, USA. [6]Michelson Science Center, California Institute of Technology, Mail Stop 100-22, Pasadena, California 91125, USA.

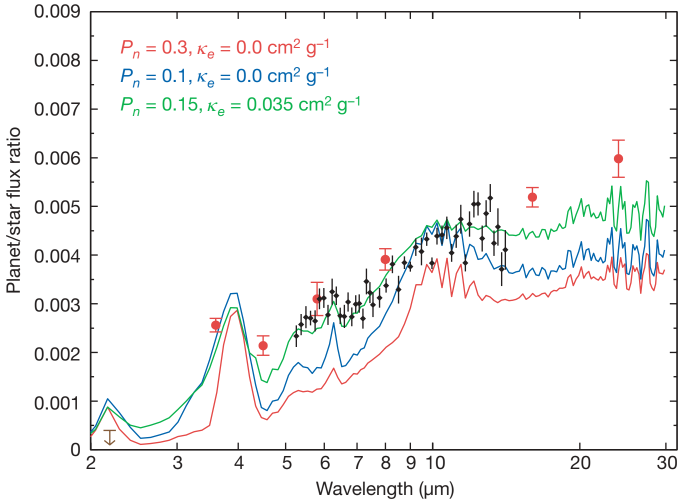


Figure 1 | Comparison of spectral observations with broadband photometry and theoretical models of the dayside atmosphere of HD 189733b. The black points show the mean planet/star flux ratios for six second-order spectra (5–8 μm) and four first-order spectra (7.5–14 μm). The data have been binned by a factor of four after light-curve fitting (corresponding to two IRS resolution elements), and the plotted uncertainties reflect the standard error in the mean in each wavelength bin. The filled red circles show broadband measurements from ref. 5 at 3.6, 4.5, 5.8, 8.0, 16 and 24 μm (error bars on this data, s.e.). The upper limit at 2.2 μm is derived from Keck spectroscopy[16]. The red, blue and green traces are atmospheric model predictions for three values of a dayside–nightside heat redistribution parameter, P_n, and two values of the extra upper-atmosphere opacity, κ_e. The model predictions have not been scaled in any way.

4.5-μm flux ratio and the decreasing planet/star flux ratio at wavelengths less than ~10 μm are signatures of the presence of gaseous water in absorption in the atmosphere of HD 189733b, with the possible presence of carbon monoxide a contributor near the IRAC 4.5-μm point[18,19]. The most interesting feature in our spectrum is the bump near ~6–6.5 μm in both the models and the data. This is the predicted[17–19] peak between the P and R branches of the ν_2 vibrational bending mode of water vapour. It appears as an emission feature because the absence of a Q branch for this mode allows for additional flux between the P and R branch absorptions.

We have investigated the significance of the 6.2-μm bump and the 7-μm trough, and a Kolmogorov–Smirnov test shows that the hypothesis that the 5.2–8-μm region of the spectrum could have been drawn from a featureless, straight-line, planet-to-star flux ratio can be rejected at the 98% confidence level. To establish the significance level of the 6.2-μm bump in a model-independent fashion, we excluded the data points from 5.9 to 6.6 μm and fitted a straight line to the remaining data between 5.2 and 8 μm. Using a chi-squared test, we find that we can reject the hypothesis that the observed data points between 5.8 and 6.6 μm are consistent with a straight line at the 99.992% probability level. A significant component of this result is contributed by an apparent peak at 5.9 μm that is not due to water vapour and is not predicted in any of the model spectra. Goodness-of-fit statistics show no indication of poor light-curve fits at or near this wavelength and we have no a priori reason to discount the data. However, if we exclude the data between 5.8 and 6.0 μm, we can still reject the null hypothesis at the 99.35% level.

Although the models without an extra absorber (that is, with $\kappa_e = 0$) reproduce all the basic features of the observations, the absolute flux levels are much better reproduced by a model with an extra upper-atmosphere absorber ($\kappa_e = 0.035\ \mathrm{cm^2\,g^{-1}}$)[17]. This model also better reproduces the shallower spectral slope between 7 and 10 μm. However, such a low value of κ_e does not produce a thermal inversion, merely a slightly hotter upper atmosphere. Unlike for HD 209458b (refs 20, 21), the ratio between the IRAC 3.6-μm and 4.5-μm points is greater than one and the IRAC 5.8-μm point is not far above a straight line between 3.6 and 4.5 μm. The former is consistent with the low (but non-zero) value of κ_e needed for the best fit or with a greater carbon monoxide abundance, whereas the latter may imply lower water abundance. Indeed, differences in the abundances, in the character and depth of heat redistribution, and in the nature and strength of a possible extra stratospheric absorber might be implicated in the emerging differences seen in the Spitzer data at secondary eclipse for the family of close-by extrasolar giant planets. Our models use a default prescription for the depth of heat circulation by super-rotational winds[17], so we suspect that a more comprehensive study that explores this dependence will further constrain the character of zonal heat transport. Moreover, performing two-dimensional spectral calculations that better incorporate the integrated slant-angle effects on the planet's dayside hemisphere may also be expected to improve the agreement.

Our spectrum is best fitted by models with relatively low heat redistribution efficiencies. The apparent conflict between this and the high nightside temperature of HD 189733b (ref. 7) and models that fit the broadband photometry has already been noted[5]. Our spectrum reinforces this conflict and, to this extent, supports suggestions[5] that the degree of heat redistribution may depend on atmospheric depth and could require three-dimensional modelling to be understood[22], or that HD 189733b has an internal energy source and radiates more energy than it receives through insolation.

Whereas the current measurements show a pronounced downturn below 10 μm, our previous IRS observations[1] of HD 189733b found the flux ratios to be essentially flat over the 7.5–14.5-μm wavelength range. Despite the smaller wavelength range, the downward slope at wavelengths less than ~10 μm should have been detectable in the earlier work. We note that a similar flat IRS spectrum was found for the giant extrasolar planet HD 209458b (refs 23, 24), although this may have been due to the thermal inversion in its atmosphere[20,21]. A separate study[25] noted that a physically implausible source of opacity would have been required to reconcile the results of the earlier IRS observations with the 8-μm photometric value[7], suggesting that the IRS systematic errors may have been larger than appreciated. Using the scatter in the present 7.5–14.5-μm spectra as a measure of the uncertainty per transit observation, a comparison of our new result with the previous spectrum of HD 189733b yields a reduced χ^2 value of 1.9, indicating that the previous result is unlikely to have been due to a random departure from the mean. Similarly, if we fit slopes to the flux ratios as a function of wavelength, we find that the slope measured for the 2006 data departs at the 3σ level from the mean slope for the current observations. Whereas the average flux ratios among the current spectra differ at the 10% level, and the slopes in the flux ratios differ at the 30% level, our previous result is clearly out of character. The only difference between the present and the earlier data was in the use of longer exposure times and access to a wider wavelength range; the analysis procedures are essentially identical and a reanalysis of the 2006 data continues to produce a flat spectrum. There are currently no known systematic differences in the relative calibration of the shorter and longer exposures taken with the IRS.

A third possibility is that the upper atmosphere of HD 189733b may actually change with time. For tidally locked, slowly rotating hot Jupiters, models predict that dynamical weather structures will form on planet-spanning scales[6]. Further observations over a longer period will be required to substantiate such a hypothesis, and several more IRS transit observations of HD 189733b are planned for 2008.

Received 13 August; accepted 17 October 2008.

1. Grillmair, C. J. *et al.* A Spitzer spectrum of the extrasolar planet HD 189733b. *Astrophys. J.* **685,** L115–L118 (2007).
2. Burrows, A. A theoretical look at the direct detection of giant planets outside the solar system. *Nature* **433,** 261–268 (2005).
3. Swain, M. E., Vasisht, G. & Tinetti, G. Methane present in an extrasolar planet atmosphere. *Nature* **452,** 329–331 (2008).
4. Tinetti, G. *et al.* Water vapour in the atmosphere of a transiting extrasolar planet. *Nature* **448,** 169–171 (2007).

5. Charbonneau, D. *et al.* The broadband infrared emission spectrum of the exoplanet HD 189733b. *Astrophys. J.* **686,** 1341–1348 (2008).
6. Rauscher, E. *et al.* On signatures of atmospheric features in thermal phase curves of hot Jupiters. *Astrophys. J.* **681,** 1646–1652 (2008).
7. Knutson, H. *et al.* A map of the day-night contrast of the extrasolar planet HD 189733b. *Nature* **447,** 183–186 (2007).
8. Bouchy, F. *et al.* ELODIE metallicity-biased search for transiting hot Jupiters. II. A very hot Jupiter transiting the bright K star HD 189733. *Astron. Astrophys.* **444,** L15–L19 (2005).
9. Charbonneau, D., Brown, T. M., Burrows, A. & Laughlin, G. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) 701–716 (Univ. Arizona Press, 2007).
10. Torres, G., Winn, J. N. & Holman, M. J. Improved parameters for extrasolar transiting planets. *Astrophys. J.* **677,** 1324–1342 (2008).
11. Ehrenreich, D. *et al.* A Spitzer search for water in the transiting exoplanet HD 189733b. *Astrophys. J.* **668,** L179–L182 (2007).
12. Pont, F., Knutson, H., Gilliland, R. L., Moutou, C. & Charbonneau, D. Detection of atmospheric haze on an extrasolar planet : the 0.55–1.05 μm transmission spectrum of HD 189733b with the Hubble Space Telescope. *Mon. Not. R. Astron. Soc.* **385,** 109–118 (2008).
13. Barman, T. On the presence of water and and global circulation in the transiting planet HD 189733b. *Astrophys. J.* **676,** L61–L64 (2008).
14. Houck, J. R. *et al.* The Infrared Spectrograph (IRS) on the Spitzer Space Telescope. *Astrophys. J. Suppl. Ser.* **154,** 18–24 (2004).
15. Deming, D. *et al.* Strong infrared emission from the extrasolar planet HD 189733b. *Astrophys. J.* **644,** 560–564 (2006).
16. Barnes, J. R. *et al.* Limits on the 2.2μm contrast ratio of the close-orbiting planet HD 189733b. *Mon. Not. R. Astron. Soc.* **382,** 473–480 (2007).
17. Burrows, A., Budaj, J. & Hubeny, I. Theoretical spectra and light curves of close-in extrasolar giant planets and comparison with data. *Astrophys. J.* **678,** 1436–1457 (2008).
18. Burrows, A., Sudarsky, D. & Hubeny, I. Theory for the secondary eclipse fluxes, spectra, atmospheres, and light curves of transiting extrasolar giant planets. *Astrophys. J.* **650,** 1140–1149 (2006).
19. Burrows, A., Hubeny, I. & Sudarsky, D. A theoretical interpretation of the measurements of the secondary eclipses of TrES-1 and HD 209458b. *Astrophys. J.* **625,** L135–L138 (2005).
20. Knutson, H. A., Charbonneau, D., Allen, L. E., Burrows, A. & Megeath, S. T. The 3.6–8.0 micron broadband spectrum of HD 209458b: Evidence for an atmospheric inversion. *Astrophys. J.* **673,** 526–531 (2008).
21. Burrows, A., Hubeny, I., Budaj, J., Knutson, H. A. & Charbonneau, D. Theoretical spectral models of the planet HD 209458b with a thermal inversion and water emission bands. *Astrophys. J.* **668,** L171–L174 (2007).
22. Showman, A. P. *et al.* Atmospheric circulation of hot Jupiters: Coupled radiative-dynamical general circulation model simulations of HD 189733b and HD 209458b. *Astrophys. J.* (submitted).
23. Richardson, L. J. *et al.* A spectrum of an extrasolar planet. *Nature* **445,** 892–895 (2007).
24. Swain, M. R. *et al.* The mid-infrared spectrum of the transiting exoplanet HD 209458b. *Astrophys. J.* **674,** 482–497 (2008).
25. Fortney, J. J. & Marley, M. S. Analysis of Spitzer spectra of irradiated planets: Evidence for water vapor? *Astrophys. J.* **666,** L45–L48 (2008).

Acknowledgements This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory (JPL), California Institute of Technology (Caltech), under a contract with NASA. Support for this work was provided by NASA through an award issued by JPL/Caltech. This study was supported in part by NASA (grant NNGO4GL22G).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.J.G. (carl@ipac.caltech.edu).

LETTERS

Strong ocean tidal flow and heating on moons of the outer planets

Robert H. Tyler[1]

Data from recent space missions have added strong support for the idea that there are liquid oceans on several moons of the outer planets, with Jupiter's moon Europa having received the most attention[1–4]. But given the extremely cold surface temperatures and meagre radiogenic heat sources of these moons, it is still unclear how these oceans remain liquid. The prevailing conjecture is that these oceans are heated by tidal forces that flex the solid moon (rock plus ice) during its eccentric orbit, and that this heat entering the ocean does not rapidly escape because of the insulating layer of ice over the ocean surface. Here, however, I describe strong tidal dissipation (and heating) in the liquid oceans; I show that a subdominant and previously unconsidered tidal force due to obliquity (axial tilt of the moon with respect to its orbital plane) has the right form and frequency to resonantly excite large-amplitude Rossby waves in these oceans. In the specific case of Europa, the minimum kinetic energy of the flow associated with this resonance (7.3×10^{18} J) is two thousand times larger than that of the flow excited by the dominant tidal forces, and dissipation of this energy seems large enough to be a primary ocean heat source.

I consider the specific case of Europa, but the general results apply equally to other moons with suspected oceans (for example Callisto, Ganymede and Titan) because the parameter regimes are similar. In previous work, potential aspects of tidal flow on Titan have been obtained using either a simplistic analytical model[5] that unjustifiably ignores Coriolis forces, or a fully numerical simulation[6] where the behavioural dependence of solutions on input parameters is harder to ascertain. Europa's tidal surface deformation, including a quasi-static ocean response, has also been previously discussed[7]. In these previous studies only the eccentricity tidal component was considered. Obliquity tidal forces are currently receiving attention as a possible means of heating extra-solar 'hot Jupiters'[8–10] but the application is formally distinct from that of the one here involving flow confined to a thin shell.

The primary distinguishing feature of the work presented here is that I consider the dynamic response of the ocean to the tidal forces, as opposed to a quasi-static equilibrium response. To illustrate the difference, consider rocking a tub of water: if the rocking is slow enough then the water is at any moment in an equilibrium configuration with the restoring force (gravity) and there is usually no need to consider the full dynamic response which includes the momentum of the waves and flow. But resolving the full dynamic response removes the low-frequency forcing assumption and allows for the situation where one may rock the tub with just the right frequency to create a resonance and amplitudes that would not have been suggested otherwise. Following this analogy, the time needed for the water to adjust is set by the speed of surface-gravity waves $c=(gh)^{1/2}$, where g is the gravitational acceleration (equal to $1.31\ \mathrm{m^2\,s^{-1}}$ for the case of Europa) and h is the water depth. In the case of Europa, it is expected that h, although less than about 200 km, is almost certainly much greater than 1 km, implying a large enough value of c for the ocean to adjust quickly to forces varying at tidal frequencies. In this case, the equilibrium solution and neglect of the dynamical response would seem justified. What is missing, however, is that in the spherical and rotating ocean, additional Coriolis forces are present that allow for a second class of oscillations (Rossby waves) including a subclass (Rossby–Haurwitz waves) that cannot be mediated by the fast gravity waves. This is because the flow associated with Rossby–Haurwitz waves is tangentially non-divergent and therefore involves no radial (up/down) motion. Without radial displacements, gravity cannot act as a restoring force in these waves and so the fast speed of the gravity waves is irrelevant. The result is that a class of natural oscillations are possible with lower frequencies that overlap those of the tidal forces. It is in this overlap that we find the resonant configuration that I discuss here.

Tidal forces on Europa are described by the gravitational potential Φ, or alternatively the equilibrium-tide surface elevation $\eta_F=\Phi/g$, which describes the ocean surface that would be obtained if ocean waves were fast enough to keep the ocean instantly adjusted to variations in the tidal forcing. With justifications described in the Methods section, including neglect of coupling with the ice sheet for these large scales[11], ocean tides on the outer moons are governed by the same equations used to calculate ocean tides on the Earth. These are the inhomogeneous Laplace tidal equations:

$$\partial_t \mathbf{u} - f\mathbf{u}\times\hat{\mathbf{r}} = -g\nabla(\eta-\eta_F) \quad (1)$$

$$\partial_t \eta + \nabla\cdot(h\mathbf{u}) = 0 \quad (2)$$

where $\mathbf{u}$ is the ocean flow velocity, η is the sea surface displacement, f is the latitude-dependent Coriolis parameter, and $\hat{\mathbf{r}}$ is the radial unit vector.

The shape of η_F is somewhat like an American football, with one tip pointing at Jupiter's centre. If η_F were steady, the flow $\mathbf{u}$ would redistribute the ocean only until the surface η had been brought to match η_F, after which there would be no further motion. But in a Europan frame the direction (and distance) to Jupiter's centre varies slightly over the tidal period, and this leads to time-dependent forcing (η_F) and flows that attempt to adjust the ocean to these time-dependent forces. There are two sources for such variations, eccentricity and obliquity, which I consider sequentially. The ocean response in each case is found by solving the above equations, but using two different methods. The case with eccentricity tidal forcing is calculated using the semi-numerical method in ref. 12, whereas for the primary result (obliquity tidal forcing) I have found an analytical solution.

The dominant tidal force (in terms of strength although not flow response) is due to Europa's eccentric orbit, and the corresponding equilibrium displacement η_F has been described previously[7]. In this case, the varying distance to Jupiter modulates the ellipsoidal shape of η_F, returning it towards a sphere as distance increases. But in an

[1]Applied Physics Laboratory and Department of Earth and Space Sciences, University of Washington, 1013 NE 40th Street, Seattle, Washington 98105, USA.

eccentric orbit, Europa keeps a face pointed not precisely towards Jupiter's centre, but rather towards the empty focus of the elliptical orbit. The result is that in a Europan frame the position of Jupiter (as well as the tip of the η_F ellipsoid pointing at it) appears to cycle diurnally through small east/west nods (librations). Using an established method[12], with $g\nabla\eta_F$ added as a forcing term in the Laplace tidal equations, I have calculated the ocean flow and surface response as shown in Fig. 1. The sea surface elevation η is very close to that of the equilibrium tide η_F (the difference is 3.9 cm) and this validates the equilibrium assumption in previous work[7]. The maximum surface displacement is 23 m and is largely independent of the value assumed for h (here chosen to be $h = 100$ km) so long as h is assumed to be much greater than the values that provide resonance for the gravity-wave modes (which can be shown to be less than a kilometre). The maximum velocity (3.0 mm s^{-1}), however, is roughly inversely proportional to h. The total potential energy of this ocean response is 2.3×10^{18} J and the kinetic energy is 4.2×10^{15} J.

Now consider the tidal forcing due to Europa's obliquity. In the Europan frame, the tip of the η_F ellipsoid now nods north/south rather than east/west, and the small angular amplitude of these diurnal excursions corresponds to Europa's obliquity angle θ_o. Although the amplitude of θ_o is uncertain, it seems that at least an orbitally forced component of $\theta_o = 0.1°$ is to be expected[13]. To allow for a plausible free component I shall consider a range up to $\theta_o = 1.0°$ in my results.

To first order in θ_o (here in radians) and second order in a/q (where a is the moon's radius and q is its orbital radius), I calculate the time-dependent tide-raising part of the obliquity tide potential to be the following (see Methods section for details):

$$\Phi = g\eta_F = \frac{3}{2}\Omega^2 a^2 \theta_o \sin\theta\cos\theta(\cos(\phi - \Omega t) + \cos(\phi + \Omega t)) \quad (3)$$

where $\Omega = 2.05 \times 10^{-5}$ s^{-1} is both the rotation rate and orbital frequency of Europa, $a = 1{,}565$ km is the radius, and θ, ϕ and t are colatitude, longitude and time, respectively. This comprises a standing wave with components propagating eastward ($\cos(\phi - \Omega t)$) and westward ($\cos(\phi + \Omega t)$); it is the latter component that is of interest because it has the right spatial form (spherical harmonic degree of two and order of one) and the right frequency ($-\Omega$) required for the resonant excitation of Rossby–Haurwitz waves[12]. The wave excited has the special property that the associated flow is tangentially non-divergent and there is no vertical flow nor surface displacement η. This allows us to derive a simple analytical solution in terms of a stream function

$$\psi = \frac{3}{2}\Omega a^2 \theta_o \sin(\theta)\cos(\phi + \Omega t) \quad (4)$$

from which the tangential flow velocity is obtained as $\mathbf{u} = \nabla \times (\psi\hat{\mathbf{r}})$, where $\hat{\mathbf{r}}$ is the radial unit vector. It is easily verified that the solution (4), together with $\eta = 0$, is an exact and unique solution to the Laplace tidal equations (1) and (2) (see Methods). This solution is assuredly, then, the correct ocean tidal response to this forcing so long as the Laplace tidal equations (and boundary conditions) apply. The solution, shown in Fig. 2, corresponds to a small-amplitude westward diurnal precession of the ocean's spin axis about that of the solid moon; the sea surface displacement is zero (as is the potential energy), the velocity amplitude is 8.6 cm s^{-1}, and the kinetic energy is 7.3×10^{18} J. The full obliquity tidal response, including the non-resonant response to the eastward-propagating part of equation (1), can be calculated using the semi-numerical method. But the response is weak and the velocity and kinetic energy numbers just given remain the same. Potential energy, however, moves from zero to 5.8×10^{15} J.

What are the consequences of this tidal energy for the evolution and maintenance of these oceans? More precisely, is dissipation of the ocean tidal kinetic energy an important source of heat for the liquid ocean? Let us assume that dissipation of the flow kinetic energy ultimately appears as an equivalent rate of heating in the ocean. The amount of heating necessary to keep the ocean liquid depends on the rate of heat loss, which itself depends on the thickness of the overlying ice cover and whether the rheology of the ice allows it to convect. The radiogenic heat flow into the ocean from the sea floor is about 8 mW m^{-2} (ref. 14), and double this amount might be provided by tidal flexing in the solid interior, giving a total of 24 mW m^{-2} (7.4×10^{11} W total). The last value is matched by the obliquity tide presented above if the tidal kinetic energy dissipates in about 32 Europan tidal cycles; this can be compared with an estimated dissipation timescale of about 10 cycles for the Earth's dominant tidal component. But note that comparison of these dissipation 'Q factors', as they are called, is not precisely meaningful because the dissipation processes in the two oceans differ, and the value we are matching is anyway only a rough estimate of the heating rate needed to keep the ocean liquid. Important differences affecting dissipation in these oceans are the following: (1) the Earth's ocean tidal flow is forced to

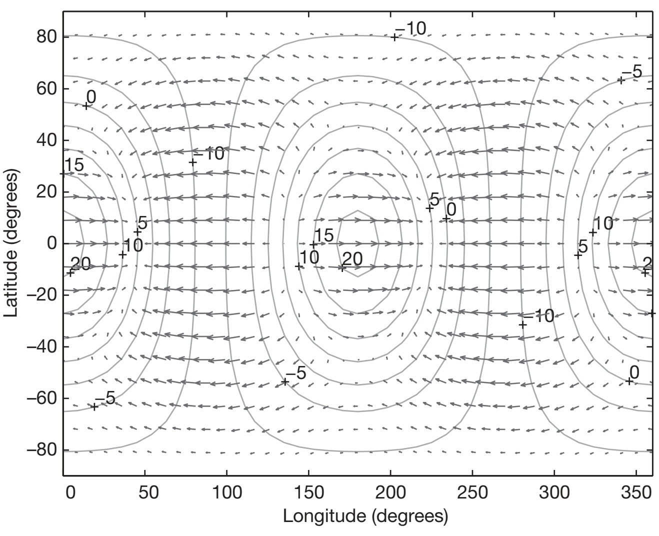


Figure 1 | Equilibrium sea surface displacement η_F (m) and tidal flow velocity vectors due to eccentricity tidal forces on Europa. This is a snapshot at phase $\omega t = 0$. Because there is nearly a quasi-static balance in this case, the sea surface η very nearly equals η_F. Primary features propagate eastward around the moon once a Europan day, with smaller standing-wave components superimposed. Flow amplitude is 3 mm s^{-1} (assuming ocean depth $h = 100$ km).

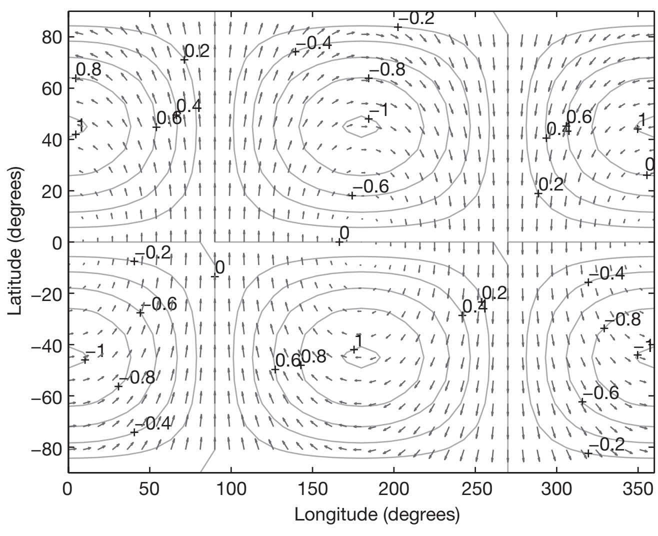


Figure 2 | Equilibrium sea surface displacement η_F (m) and tidal flow velocity vectors due to obliquity tidal forces on Europa. This is a snapshot at phase $\omega t = 0$. In this resonant case, $\eta = 0$ and all energy is in the flow field. The pattern of arrows propagates westward around the moon once a Europan day, and the contours of η_F oscillate in north/south pairs. Flow amplitude is 8.6 cm s^{-1} (for obliquity of 0.1°).

Table 1 | Ocean velocity amplitudes and minimum dissipation due to obliquity tidal forces

Obliquity (deg.)	Flow amplitude ($cm\,s^{-1}$)	Minimum dissipation (W)	Maximum dissipation (W)
0.1	8.6	6.5×10^{10}	7.9×10^{12}
0.3	26	1.9×10^{12}	7.2×10^{13}
0.5	42	8.0×10^{12}	2.0×10^{14}
0.7	60	2.2×10^{14}	3.9×10^{14}
1.0	84	6.5×10^{14}	8.0×10^{14}

Results show ocean velocity amplitudes and minimum dissipation due to obliquity tidal forces (heating) for a range of obliquity angles starting at 0.1° (the minimum orbitally forced component) and extending to 1.0° to allow for free variations. An estimate for actual dissipation allows for augmentation of dissipation by topography and internal waves but should fall below the maximum values listed or else the tidal flow response is highly damped. For reference, radiogenic heating ($8.0\,mW\,m^{-2}$) is 2.5×10^{11} W when globally integrated.

channel around continents, whereas Europa is thought to have a continuous deep ocean; (2) Europa is covered with ice and therefore has dissipative boundary layers on both the bottom and top; (3) in the Earth's deep ocean (probably the closest familiar analogy for Europa's ocean) dissipation strongly involves transference of energy to internal waves, but this transference depends on the ocean's stratification which is unknown for Europa; (4) tidal heating by flexure of Europa's ice sheet is poorly constrained but is thought to be significant[15].

Given these uncertainties, we can only reliably estimate a minimum value for the ocean tidal dissipation in Europa. From traditional boundary-layer theory, we can expect that the dissipation per unit area will equal $\rho C_d u^3$, where ρ is water density, $C_d = 0.003$ is the drag coefficient and u is the amplitude of the spatially varying flow velocity[16]. (This approach was also followed in work concerning Titan's ocean tidal dissipation[5,6].) Inserting the above flow results and integrating over the moon, we obtain the total ocean tidal dissipation. But the dissipation will be underestimated (perhaps severely) for two reasons. First, the velocities from this analytical solution are unrealistically smooth and would only be realized if the bottom of the ice layer and the sea floor were both without topographic variations. This artificial smoothness in the assumed flow decreases the integral of $\rho C_d u^3$. Second, in the Earth's deep ocean, internal waves increase dissipation over the simple boundary-layer value by a factor of about 40 (ref. 17). Taken together, a 'best guess' at the real dissipation is plausibly a factor of 80 higher than the minimum values. But I set an upper bound for the dissipation by requiring that the 'quality factor' Q of the system satisfy $Q > 6\pi$ when the water depth $h = 100$ km. For higher dissipation the response is highly damped and becomes inconsistent with assumptions used here. These maximum values are shown together with the minimum values in Table 1. The primary result in Table 1 is that even a small fraction of a degree of obliquity leads to significant dissipation and heating in the ocean. Above an obliquity angle of 0.16°, even the minimum ocean tidal dissipation surpasses the moon's radiogenic heating, and the best-guess dissipation is already over 20 times the radiogenic source at the minimum obliquity of 0.1°. These results indicate that resonant tidal flow must be considered as a significant and potentially dominant source of ocean heat on the outer moons.

METHODS SUMMARY

The general aspects of the tidal response are characterized by the 'Lamb parameter' ε. For Europa, $\varepsilon \ll 1$, indicating strong separation between Rossby and gravity-wave modes and providing easier determination of the criteria necessary for resonances in these two separated flow classes. These criteria, as well as the semi-numerical method used to obtain the flow response in the first case (eccentricity tide), have been described previously[12].

In the second case, I find an analytical solution for resonant flow under tidal forces associated with the moon's obliquity. I calculate the time-dependent tide-raising part of the obliquity tide potential to be the expression in equation (3) (see Methods for details). Inserting this expression for the forcing η_F into the governing Laplace tidal equations (1) and (2) and solving analytically, I obtain the solution as described in equation (4).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 January; accepted 16 October 2008.

1. Carr, M. H. *et al.* Evidence for a subsurface ocean on Europa. *Nature* **391,** 363–365 (1998).
2. Kivelson, M. G. *et al.* Galileo magnetometer measurements: A stronger case for a subsurface ocean at Europa. *Science* **289,** 1340–1343 (1981).
3. Stevenson, D. Europa's ocean—the case strengthens. *Science* **289,** 1305–1306 (2000).
4. Spohn, T. & Schubert, G. Oceans in the icy Galilean satellites of Jupiter? *Icarus* **161,** 456–467 (2003).
5. Sagan, C. & Dermott, F. The tide in the seas of Titan. *Nature* **300,** 731–733 (1982).
6. Sears, W. D. Tidal dissipation in oceans on Titan. *Icarus* **113,** 39–56 (1995).
7. Moore, W. B. & Schubert, G. The tidal response of Europa. *Icarus* **147,** 317–319 (2000).
8. Fabrycky, D. C., Johnson, E. T. & Goodman, J. Cassini states with dissipation: Why obliquity tides cannot inflate hot Jupiters. *Astrophys. J.* **665,** 754–766 (2007).
9. Levrard, B. *et al.* Tidal dissipation within hot Jupiters: a new appraisal. *Astron. Astrophys.* **462,** L5–L8 (2007).
10. Winn, J. N. & Holman, M. J. Obliquity tides on hot Jupiters. *Astrophys. J.* **628,** L159–L162 (2005).
11. Wadhams, P. Attenuation of swell by sea ice. *J. Geophys. Res.* **78,** 3552–3563 (1973).
12. Longuet-Higgins, M. The eigenfunctions of Laplace's tidal equations over a sphere. *Phil. Trans. R. Soc. Lond. A.* **262,** 511–607 (1968).
13. Bills, B. G. Free and forced obliquities of the Galilean satellites of Jupiter. *Icarus* **175,** 233–247 (2005).
14. Squyres, S. W., Reynolds, R. T., Cassen, P. M. & Peale, S. J. The evolution of Enceladus. *Icarus* **53,** 319–331 (1983).
15. Ojakangas, G. W. & Stevenson, D. J. Episodic volcanism of tidally heated satellites with application to Io. *Icarus* **66,** 9459–9470 (1986).
16. Egbert, G. D. & Ray, R. D. Estimates of M_2 tidal energy dissipation from TOPEX/Poseidon altimeter data. *J. Geophys. Res.* **106,** 22475–22502 (2001).
17. Jayne, S. R. & Laurent, L. C. S. Parameterizing tidal dissipation over rough topography. *Geophys. Res. Lett.* **28,** 811–814 (2001).

Acknowledgements K. Khurana provided helpful comments during the preparation of this manuscript. This work has been supported by the NASA Outer Planets Research program.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.H.T. (tyler@apl.washington.edu).

METHODS

The ocean response to tidal forces depends strongly on the propagation and dispersion of both long surface-gravity waves (ocean swell is a familiar example) and planetary-scale Rossby waves (commonly seen in global-scale weather maps). Although the presence of thick ice cover will undoubtedly modify the propagation properties of short gravity waves, extrapolation from ocean ice models suggests that the flexural and mass-loading effects due to the ice cover will be insignificant for the very long waves that I consider here[11]. More importantly, the flow in the Rossby-wave solution I present is tangentially non-divergent and there are neither the vertical flow velocities nor the normal forces that one would expect to be necessary for the ice layer to affect these waves. The general aspects of the tidal response are then, as in the ice-free case, characterized by the 'Lamb parameter' $\varepsilon = (2\Omega a/c)^2$, where $\Omega = 2.05 \times 10^{-5}\,\mathrm{s}^{-1}$ is both the rotation rate and orbital frequency of Europa, $a = 1{,}565$ km is the radius, and c as defined above is also called the shallow-water or long surface-gravity wave speed. Hence, the Lamb parameter is a squared ratio comparing the speed associated with the moon's rotation with c. For Europa, $\varepsilon \ll 1$, unlike the case for the Earth's ocean ($\varepsilon \approx 80$). Physically, this indicates that there will be a strong separation between Rossby and gravity-wave modes, with little of the mixed 'rotational-gravity' types that prevail on the Earth. Computationally, small ε simplifies analyses of the governing Laplace tidal equations and provides easier determination of the criteria necessary for resonances in these two separated flow classes. These criteria, as well as the semi-numerical method used to obtain the flow response in the first case (eccentricity tide), have been previously described in detail[12].

In the second case (the primary result in this research) I find an analytical solution for resonant flow under tidal forces associated with the moon's obliquity. In this case, the forcing η_F has not been previously described, and I have calculated it as follows. The gravitational potential at a point on Europa's surface is given by $\Phi = g\eta = -GM_J/d$, where G is the gravitational constant, M_J is Jupiter's mass and d is the distance between the point on Europa and Jupiter's centre. This distance is $d = (a^2 + q^2 - 2aq\cos\alpha)^{1/2}$, where a is Europa's radius, q is the orbital radius, and α describes the angle between the line connecting the centres of Europa and Jupiter and the line connecting Europa's centre with the point on its surface that we are considering. Considering that a and q are constants in this case we calculate $\cos\alpha = \cos\tilde{\theta}\sin\theta\cos\phi + \sin\tilde{\theta}\cos\theta$, where θ and ϕ are the colatitude and longitude (measured from sub-Jovian point) of the Europa-fixed coordinate frame, and $\tilde{\theta} = \theta_o\cos(\Omega t)$ describes the small north/south variations in the apparent position of Jupiter. To first order in θ_o (here in radians) and second order in a/q (where a is the moon's radius and q is its orbital radius), I calculate the time-dependent tide-raising part of the obliquity tide potential to be the expression in equation (3). Inserting this expression for η_F into the governing Laplace tidal equations (1) and (2) and solving analytically, we obtain the solution as described in equation (4).

Vol 456 | 11 December 2008 | doi:10.1038/nature07607

A solid-state light–matter interface at the single-photon level

Hugues de Riedmatten[1]*, Mikael Afzelius[1]*, Matthias U. Staudt[1], Christoph Simon[1] & Nicolas Gisin[1]

Coherent and reversible mapping of quantum information between light and matter is an important experimental challenge in quantum information science. In particular, it is an essential requirement for the implementation of quantum networks and quantum repeaters[1–3]. So far, quantum interfaces between light and atoms have been demonstrated with atomic gases[4–9], and with single trapped atoms in cavities[10]. Here we demonstrate the coherent and reversible mapping of a light field with less than one photon per pulse onto an ensemble of ~10^7 atoms naturally trapped in a solid. This is achieved by coherently absorbing the light field in a suitably prepared solid-state atomic medium[11]. The state of the light is mapped onto collective atomic excitations at an optical transition and stored for a pre-determined time of up to 1 μs before being released in a well-defined spatio-temporal mode as a result of a collective interference. The coherence of the process is verified by performing an interference experiment with two stored weak pulses with a variable phase relation. Visibilities of more than 95 per cent are obtained, demonstrating the high coherence of the mapping process at the single-photon level. In addition, we show experimentally that our interface makes it possible to store and retrieve light fields in multiple temporal modes. Our results open the way to multimode solid-state quantum memories as a promising alternative to atomic gases.

Efficient and reversible mapping of quantum states between light and matter requires strong atom–photon interaction. This can be achieved using ensembles of atoms, where light can be efficiently absorbed and where it is possible to engineer the atomic systems such that the stored light can be retrieved in a well-defined spatio-temporal mode owing to there being a collective constructive interference between all the emitters. This collective enhancement is at the heart of protocols for storing photonic quantum states in atomic ensembles, such as schemes based on electromagnetically induced transparency (EIT)[12], off-resonant Raman interactions[13] and modified photon echoes using controlled reversible inhomogeneous broadening (CRIB)[14,15] and atomic frequency combs (AFC)[11].

Certain solid-state systems have properties that make them very attractive for applications in quantum state storage. In particular, solids doped with rare-earth ions provide a unique physical system in which large ensembles of atoms are naturally trapped in a solid-state matrix, which prevents decoherence due to the motion of the atoms. These systems exhibit excellent coherence properties at low temperature (below 4 K), both for the optical[16] and the spin transitions[17]. Classical light has been stored for more than 1 s in a rare-earth-doped solid[18]. The long optical coherence times enable storage of multiple temporal modes in a single quantum memory, which promises significant speed increases in quantum repeater applications[3]. Furthermore, the high optical densities required to achieve high-efficiency light storage and retrieval can be obtained. However, despite recent experimental progress[18–23], the implementation of a solid-state light–matter quantum interface has not yet been reported.

Here we demonstrate the coherent mapping of light at the single-photon level onto a large number of atoms in a solid, and collective re-emission of the stored light at a pre-determined time. Time-delayed interferometry with weak X-ray pulses scattered by collective states of iron nuclei was demonstrated in this context in ref. 24. In our experiment in the optical regime, the mapping is done by coherently absorbing the light in an ensemble of inhomogeneously broadened atoms spectrally prepared with a periodic modulation of the absorption profile[11,25]. The reversible absorption by such a spectral grating is at the heart of the recently proposed multimode quantum memory scheme based on AFC[11]. We therefore also demonstrate a proof of principle of an essential primitive of this protocol at the single-photon level.

Let us now describe how our interface works. The inhomogeneous spectral atomic distribution at the optical transition between the ground state, $|g\rangle$, and the excited state, $|e\rangle$, is shaped into a series of absorbing peaks of width γ, with a peak separation Δ (Fig. 1a). This can be done using optical pumping techniques, as discussed later. Once the spectral grating is prepared, we send in weak coherent states of light, $|\alpha\rangle_L$, with mean photon number $\bar{n} = |\alpha|^2 < 1$. The bandwidth of each photon must be larger than Δ but smaller than the total width of the grating. It is then possible to show that the light can be absorbed uniformly over its entire frequency spectrum, although the atomic spectral distribution has large gaps[11]. This can be understood from the time–energy Heisenberg uncertainty relations. Over the short timescale given by the duration, τ_P, of the photon, the linewidth of the atomic resonance is effectively broadened to a value of order $1/\tau_P > \Delta$, resulting in a uniform, but smaller, absorption. The effective optical depth of the spectral grating is given by $\tilde{d} \approx d/F$, where $F = \Delta/\gamma$ is the finesse of the spectral grating and d the grating optical depth.

After absorption, the light field is stored in a coherent superposition of collective optical excitations delocalized over all the resonant atoms in the grating. The state of the atoms (not normalized) can be written as

$$|\alpha\rangle_A = |0\rangle_A + \alpha|1\rangle_A + O(\alpha^2)$$

where $|0\rangle_A = |g_1 \ldots g_N\rangle$ and

$$|1\rangle_A = \sum_j c_j e^{i\delta_j t} e^{ikz_j} |g_1 \cdots e_j \cdots g_N\rangle$$

Here N is the total number of atoms; $|g_j\rangle$ and $|e_j\rangle$ represent the ground and excited states, respectively, of atom j; z_j is the position of atom j; k is the wavenumber of the light field (for simplicity, we only consider the forward-propagating spatial mode); δ_j is the detuning of the atom with respect to the laser frequency; and the amplitudes c_j depend on

[1]Group of Applied Physics, University of Geneva, CH-1211 Geneva 4, Switzerland.
*These authors contributed equally to this work.

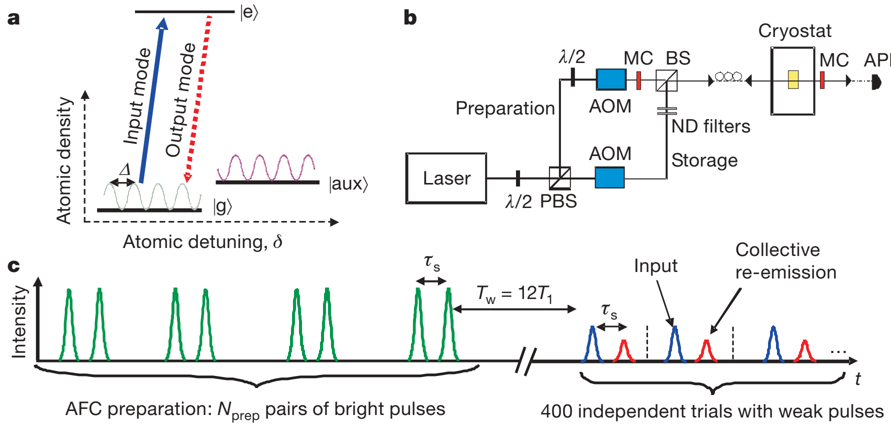


Figure 1 | Overview of the experiment. a, Solid-state light–matter interface. The light at 880 nm is absorbed at the $^4I_{9/2} \rightarrow {}^4F_{3/2}$ transition of Nd^{3+} ions, doped into a YVO_4 crystal (doping level, 10 p.p.m.). The inhomogeneous broadening of the optical transition is 2 GHz, with a maximal optical depth of around four for a 1-mm-long sample. The optical relaxation time, T_1, of the excited state $|e\rangle$ is equal to 100 μs. The ground state is split into two Zeeman levels, $|g\rangle$ and $|aux\rangle$, separated by 3.8 GHz through the application of a magnetic field of 300 mT applied along the crystal *c* axis. A spectral grating is prepared in $|g\rangle$ following a preparation sequence described in Methods. **b**, Experimental setup. The laser source is a continuous-wave, external-cavity diode laser at 880 nm. The laser is split at a variable beam splitter made using a half-wave plate and a polarising beam splitter (PBS). The pulse sequences for the preparation of the grating and for the pulses to be stored are prepared by independent acousto-optic modulators (AOM) in the different optical paths. Neutral density (ND) filters are used to decrease the intensity of the pulses to be stored to the single-photon level. The duration of the pulses is about 30 ns. These paths are then recombined at a beam splitter (BS) and coupled into a single mode fibre to ensure proper mode matching. The light is then focused onto the crystal with a beam diameter of 30 μm. The crystal is cooled to 3 K using a pulse tube cooler. Beyond the sample, the light passes through a polarizer and is coupled back to a single mode fibre, which is connected to a silicon avalanche-photodiode (APD) single-photon counter. To block the preparation light during storage and to protect the detector from the intense preparation light, two mechanical choppers (MCs) are used. **c**, Optical pulse sequence. The experimental sequence is divided into two parts: the preparation of the spectral grating and the storage of the weak pulses. The preparation is done with a sequence of N_{prep} pairs of pulses (see Methods). We wait a time $T_w = 12T_1 = 1{,}200$ μs between the preparation and the storage sequence to avoid fluorescence noise. During the storage sequence, 400 independent trials are performed at a repetition rate of 200 kHz. The entire sequence preparation plus storage is then repeated with a repetition rate of 40 Hz.

the frequency and on the spatial position of atom *j*. This collective state will rapidly dephase because each term acquires an individual phase $e^{i\delta_j t}$. For narrow absorption peaks, the detunings can be written $\delta_j = m_j\Delta$, where m_j are integers. Owing to this periodic structure of the absorption profile, the collective state will then be re-established after a pre-determined time $2\pi/\Delta$. This leads to a coherent photon-echo-type re-emission in the forward spatial mode[11,25]. If the absorption peaks have a finite width, the rephasing will not be perfect, leading to a reduction of the collective signal[11], as discussed below and in the Supplementary Information. We note that in our experiment the light field is stored as a collective excitation at the optical transition, contrary to all previous light storage experiments at the single-photon level, where collective excitations of spin states were used[5–7].

The solid-state interface is implemented in an ensemble of neodymium ions (Nd^{3+}) doped into a YVO_4 crystal[26]. The Nd^{3+} ions constitute an ensemble of inhomogeneously broadened atoms having a relevant level structure with two spin ground states, $|g\rangle$ and $|aux\rangle$, and one excited state, $|e\rangle$, as shown in Fig. 1. Initially, the two ground states are equally populated for all frequencies within the inhomogeneous broadening. The preparation of the spectral grating is achieved by frequency-selective optical pumping from $|g\rangle$ to $|aux\rangle$ through the excited state $|e\rangle$ (see Fig. 1 for an overview of the experiment). This is implemented with a Ramsey-type interference using a train of coherent pairs of pulses[27] (Fig. 1, Methods). To store weak light fields, it is required that there be no population in the excited state, as otherwise fluorescence will blur the signal. This is ensured by waiting long enough between the preparation and storage sequences for all atoms to return to the ground states. As a result of the preparation sequence, a spectral grating is present in $|g\rangle$ before the storage begins. The grating decays with the population relaxation lifetime T_Z between the spin states ($T_Z = 6$ ms in our case).

In the first experiment, we demonstrate collective mapping of weak coherent states of light, $|\alpha\rangle_L$, onto the crystal. An example with $\bar{n} = 0.5$ is shown in Fig. 2 (See Methods for the estimation of $\bar{n}$). When the sample is prepared with a spectral grating having a periodicity of 4 MHz, we observe a strong emission at the expected storage time of 250 ns. The measured storage and retrieval efficiency is $\eta \approx 0.005$, which means that 0.5% of the incoming light is re-emitted in this signal. Although η is low, the re-emission probability is more than four orders of magnitude larger than what would be expected from a non-collective spontaneous re-emission, taking into account that we collect a solid angle of 2×10^{-4} and that the optical relaxation time is three orders of magnitude greater than the observed signal time. This signal thus clearly arises from a collective re-emission, which

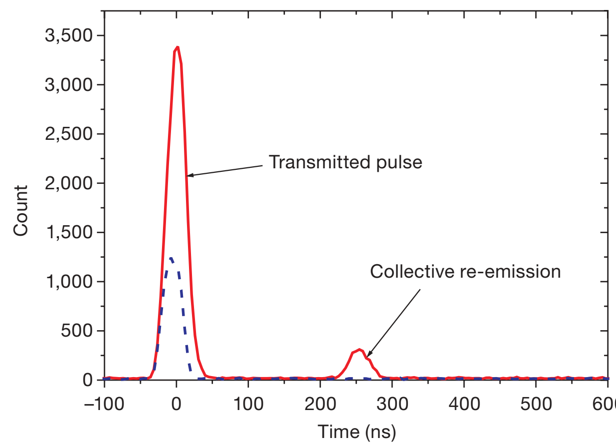


Figure 2 | Reversible mapping of a coherent state with $\bar{n} = 0.5$. The solid line corresponds to the case in which a spectral grating is prepared with a periodicity of 4 MHz. The peak 250 ns after the transmitted input pulse corresponds to the collective retrieval after storage in the solid-state medium. The dashed line corresponds to the case in which the atomic medium is not prepared with a spectral grating. In this case, we only see the transmitted input pulse (about 2% of the incoming pulse is transmitted). The absorption of the input pulse is smaller when the spectral grating is prepared (about 5% of the light is transmitted).

demonstrates the collective and reversible mapping of a light field with less than one photon on average onto a large number of atoms in a solid. To further study the mapping process, we record the number of counts in the observed signal for values of $\bar{n}$ ranging from 0.2 to 2.7 (Fig. 3a). This shows that the mapping is linear and that very low photon numbers can still be mapped and retrieved. We also investigated the decay of storage efficiency with storage time (Fig. 3b).

So far we have considered the storage of weak light fields in a single temporal mode. However, the use of spectral gratings also allows for storage in multiple temporal modes, as pointed out in ref. 11. To illustrate this multimode property, we store trains of four weak pulses, with values of $\bar{n}$ ranging from 0.8 to 0.3, for 500 ns, as shown in Fig. 4. The maximal number of modes that can be stored is given by the ratio of the storage time (determined by the spectral grating) to the duration of an individual mode. In the present experiment, the shortest duration of pulses was set to about 20 ns (full-width at half-maximum) by technical limitations. A great advantage of the AFC protocol is that the number of modes that can be stored does not depend on the optical depth, unlike for EIT and CRIB[11].

For applications in quantum memories, it is crucial that the interface conserves the phase of the incoming pulses. To probe the coherence of the mapping process, we store a pair of weak pulses that are separate by a time $\tau = 100$ ns and have a variable relative phase, Φ. This can be viewed as a time-bin qubit, which can be written as $|\psi\rangle_L = |\alpha_t\rangle_L + e^{i\Phi}|\alpha_{t+\tau}\rangle_L$, where $|\alpha_t\rangle_L$ represents a weak coherent state of light at time t. The qubit is stored and thereafter analysed directly in the memory[21]. This method requires the implementation of partial readouts at different times. This realizes a projection on a superposition basis, similar to what can be done with an unbalanced Mach–Zehnder interferometer[21]. If the time, τ, between the weak pulses matches the time between the two readouts, the re-emission from the sample can be suppressed or enhanced depending on the phase, Φ. The visibility of this interference is a measure of the coherence of the mapping process. The two partial readouts are made by preparing two superimposed spectral gratings of different periods, corresponding to respective storage times of 200 and 300 ns. An example of interference for $\bar{n} = 0.85$ is shown in Fig. 5. We measured net visibilities (that is, after subtraction of detector dark counts) above 95% for various values of $\bar{n}$ between 0.4 and 1.7, which demonstrates the high coherence of the storage process, even at the single-photon level. This excellent phase preservation results from the collective enhancement[22] and the almost complete suppression of background noise.

We now analyse in more detail the efficiency performance of our interface using the model developed for the AFC quantum memory[11]. The efficiency of the storage and retrieval in the forward mode, for an ideal AFC, is given by (Supplementary Information)

$$\eta_{\mathrm{AFC}} \approx (d/F)^2 e^{-7/F^2} e^{-d/F} \qquad (1)$$

The first factor represents the collective interaction, which is proportional to the square of the number of atoms in the spectral grating.

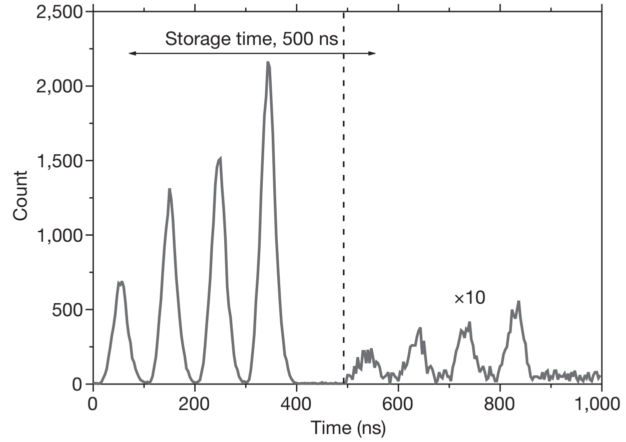


Figure 4 | Multimode light–matter interface. The spectral grating is prepared for a storage time of 500 ns. The first four pulses are the transmitted pulses. After 500 ns, we can clearly see the collective re-emission of the four temporal modes. The signal to noise ratio is smaller than for the single-mode case (Fig. 2), owing to the longer storage time of 500 ns (compare with Fig. 3b). For clarity, the output signal part has been magnified by a factor of ten.

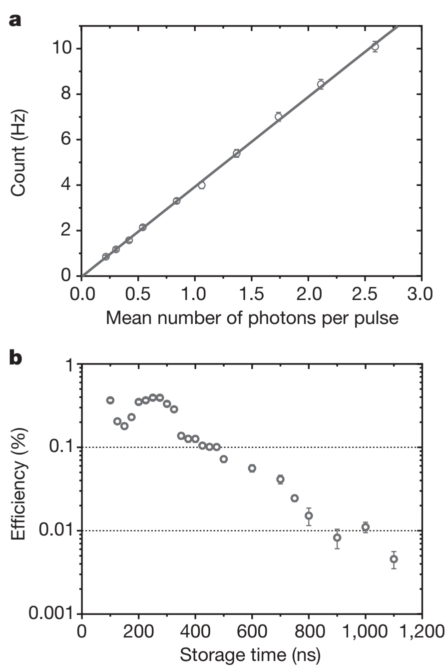


Figure 3 | Study of the mapping process. a, Number of detections in the collective output mode as a function of $\bar{n}$. **b**, Efficiency as a function of the storage time for $\bar{n} = 2.7$. For small storage times (<400 ns), an oscillation in the efficiency is clearly visible. This is a quantum beat due to the interaction of the electronic spin of the Nd^{3+} ion with the nuclear spin of the surrounding vanadium ions (called a super-hyperfine interaction). For longer storage times, the decay is exponential with decay constant of 220 ns. This interaction also limits the minimal width of the absorption peaks. Error bars represent statistical uncertainties in photon count (±1 s.d.).

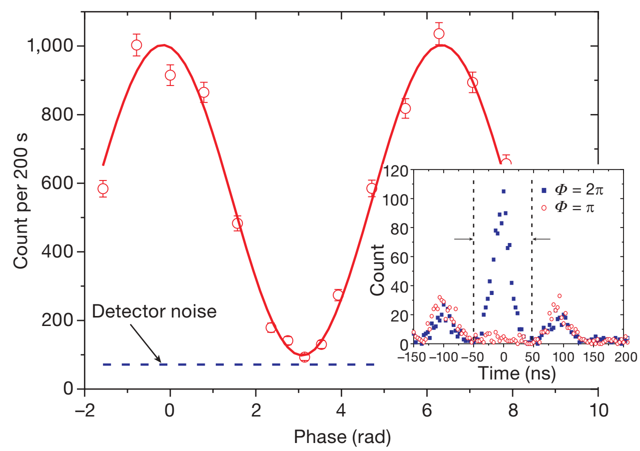


Figure 5 | Interference fringes. Time-bin qubits with different phases, Φ, are stored and analysed using the interface. The analysis is performed by projecting the time-bin qubit on a fixed superposition basis, which here is achieved by two partial readouts (see text for details). The inset shows the histogram of arrival times, where there is constructive interference for $\Phi = 2\pi$ and destructive interference for $\Phi = \pi$ in the middle time bin. For this particular interference fringe, we obtain a raw visibility of 82% and a visibility of 95% when subtracting detector dark counts. Errors bars represent statistical uncertainties of photon counts (±1 s.d.).

The second factor describes the intrinsic AFC dephasing during the storage due to the finite width of the absorption peaks. The last factor accounts for the re-absorption of the emitted photon by the resonant atoms. In the present experiment, imperfect optical pumping results in a uniform absorbing background with residual optical depth d_0 that will act as a passive loss. The efficiency is then given by $\eta = \eta_{\mathrm{AFC}} e^{-d_0}$ (Supplementary Information). Using the model presented in the Supplementary Information, we estimate that $F \approx 2$, $d = 1.44 \pm 0.08$ and $d_0 = 2.28 \pm 0.14$. This large value of d_0 means that 90% of incoming photons are lost to the background atoms, which do not give rise to a collective emission.

To analyse the physics of the collective enhancement further, it is informative to infer the intrinsic storage efficiency, η_{AFC}, that could be obtained in the absence of an absorbing background. Using equation (1) and the values of d and F, we find that $\eta_{\mathrm{AFC}} = 0.04$. The intrinsic storage efficiency can be further decomposed into two terms, the write efficiency (η_{w}) and the read efficiency (η_{r}), such that $\eta_{\mathrm{AFC}} = \eta_{\mathrm{w}}\eta_{\mathrm{r}}$. The write efficiency is defined as the probability that a photon is coherently absorbed by the atoms of the AFC: $\eta_{\mathrm{w}} = 1 - e^{-d/F} = 0.51$. This implies that the read efficiency, that is, the probability that the photon is re-emitted in the collective mode conditioned on a coherent absorption, is $\eta_{\mathrm{r}} = 0.085$. To increase the write efficiency, we must increase the optical depth. The read efficiency is mostly limited by the intrinsic dephasing during the storage, owing to the low finesse, $F \approx 2$. In our experiment, F is limited by the minimal width of the absorption peak, which is of order 1–2 MHz owing to the linewidth of our free-running laser and to material properties (Fig. 3).

We emphasize, however, that these are not fundamental limitations of rare-earth-doped materials. The preparation of narrow absorption lines with a low absorption background has been demonstrated for several other rare-earth-doped materials, such as Eu^{3+}:Y_2SiO_5 (ref. 19), Pr^{3+}:Y_2SiO_5 (refs 20,28) and Tm^{3+}:YAG (ref. 29). Significantly higher storage efficiencies should be obtained in these materials. Ultimately, the efficiency of the forward retrieval is limited to 54%, owing to the re-absorption of the collective signal[11]. This can be overcome by using backward retrieval; see below.

To use our interface in a quantum repeater architecture, several further steps must be undertaken. The photon to be stored must be correlated with another photon that will be sent over long distances[3]. Moreover, on-demand readout of the stored field and longer storage times are required. This is possible by transferring the excitations in $|e\rangle$ to another long-lived, ground-state spin level, $|s\rangle$, using an optical control pulse[11], as for other quantum memory proposals, such as EIT or Raman-based schemes[30]. For on-demand readout, the excitation can be transferred back to $|e\rangle$ using another control pulse at a chosen time. Efficient and coherent transfer in an optical Λ configuration has been demonstrated in praseodymium-doped solids[28], which possess the necessary level structure of three ground-state levels ($|g\rangle$, $|\mathrm{aux}\rangle$ and $|s\rangle$). For neodymium, hyperfine states may also in principle be used. The pair of optical control pulses, if applied in a counter-propagating fashion, also change the phase pattern of the stored excitation, such that the collective field will be re-emitted backward. Then the storage efficiency can approach unity, owing to a constructive interference that effectively suppresses re-absorption[11]

We have demonstrated the collective and reversible mapping of a light field with less than one photon on average onto a large number of atoms embedded in a solid. This represents the observation of collective enhancement at the single-photon level in a solid in the optical regime. We have also demonstrated that the quantum coherence of incident weak light fields is almost perfectly conserved during the storage. Solid-state systems can therefore be considered as a promising alternative to atomic gases for photonic quantum storage. This line of research holds promise for the implementation of efficient, long-distance quantum networks.

METHODS SUMMARY

The spectral grating is prepared using a series of pulse pairs of pulse area $\theta < \pi/2$ resonant with the $|g\rangle \rightarrow |e\rangle$ transition. Each pair of pulses produces a frequency-selective coherent transfer of population from $|g\rangle$ to $|e\rangle$[27]. This can be seen as a Ramsey-type interference in which the two light pulses act as beam splitters and the phase shift acquired in the excited state depends on the detuning of the atoms. The periodicity of the created spectral grating is then given by the inverse of the time interval, τ_s, between the two pulses. The atoms in the excited state can decay to both ground states, $|g\rangle$ and $|\mathrm{aux}\rangle$, with a relaxation time of $T_1 = 100\,\mu s$. The atoms that decay to $|\mathrm{aux}\rangle$ are not affected by the preparation laser and remain in this state for a time $T_Z = 6$ ms (ref. 26). The pulse sequence is then repeated 100 times with a time separation between the pairs of 15 μs, which is longer than the optical coherence time, $T_2 = 7\,\mu s$. This allows for the build-up of the spectral grating, with population storage in $|\mathrm{aux}\rangle$. For the interference experiment, the two gratings corresponding to storage times of 200 and 300 ns are created in a similar way. Each grating is prepared with a pulse pair having the corresponding time separation, sent alternately with a time separation of 7.5 μs.

To estimate the mean number of photons per pulse, we shift the laser out of resonance with the absorbing atoms and record the proportion of detections in the single-photon counter (typically between 1% and 20%). By a careful measurement both of the detection efficiency ($\eta_{\mathrm{d}} = 0.32$) and of the transmission efficiency from the input face of the cryostat to the detector (typically $\eta_{\mathrm{t}} = 0.2$), we can infer the mean number of photons in front of the cryostat, before the sample.

Received 27 June; accepted 30 October 2008.

1. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
2. Duan, L.-M., Lukin, M. D., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
3. Simon, C. *et al.* Quantum repeaters with photon pair sources and multimode memories. *Phys. Rev. Lett.* **98**, 190503 (2007).
4. Julsgaard, B., Sherson, J., Cirac, J. I., Fiurášek, J. & Polzik, E. S. Experimental demonstration of quantum memory for light. *Nature* **432**, 482–486 (2004).
5. Chanelière, T. *et al.* Storage and retrieval of single photons transmitted between remote quantum memories. *Nature* **438**, 833–836 (2005).
6. Eisaman, M. D. *et al.* Electromagnetically induced transparency with tunable single-photon pulses. *Nature* **438**, 837–841 (2005).
7. Chou, C. W. *et al.* Measurement-induced entanglement for excitation stored in remote atomic ensembles. *Nature* **438**, 828–832 (2005).
8. Honda, K. *et al.* Storage and retrieval of a squeezed vacuum. *Phys. Rev. Lett.* **100**, 093601 (2008).
9. Appel, J., Figueroa, E., Korystov, D., Lobino, M. & Lvovsky, A. I. Quantum memory for squeezed light. *Phys. Rev. Lett.* **100**, 093602 (2008).
10. Boozer, A. D., Boca, A., Miller, R., Northup, T. E. & Kimble, H. J. Reversible state transfer between light and a single trapped atom. *Phys. Rev. Lett.* **98**, 193601 (2007).
11. Afzelius, M., Simon, C., de Riedmatten, H. & Gisin, N. Multimode quantum memory based on atomic frequency combs. Preprint at ⟨http://arxiv.org/abs/0805.4164⟩ (2008).
12. Fleischhauer, M. & Lukin, M. D. Dark-state polaritons in electromagnetically induced transparency. *Phys. Rev. Lett.* **84**, 5094–5097 (2000).
13. Nunn, J. *et al.* Mapping broadband single-photon wave packets into an atomic memory. *Phys. Rev. A* **75**, 011401 (2007).
14. Moiseev, S. A. & Kröll, S. Complete reconstruction of the quantum state of a single-photon wave packet absorbed by a Doppler-broadened transition. *Phys. Rev. Lett.* **87**, 173601 (2001).
15. Kraus, B. *et al.* Quantum memory for nonstationary light fields based on controlled reversible inhomogeneous broadening. *Phys. Rev. A* **73**, 020302 (2006).
16. Böttger, T., Thiel, C. W., Sun, Y. & Cone, R. L. Optical decoherence and spectral diffusion at 1.5 μm in $Er^{3+}Y_2SiO_5$ versus magnetic field, temperature, and Er^{3+} concentration. *Phys. Rev. B* **73**, 075101 (2006).
17. Fraval, E., Sellars, M. J. & Longdell, J. J. Dynamic decoherence control of a solid-state nuclear-quadrupole qubit. *Phys. Rev. Lett.* **95**, 030506 (2005).
18. Longdell, J. J., Fraval, E., Sellars, M. J. & Manson, N. B. Stopped light with storage times greater than one second using electromagnetically induced transparency in a solid. *Phys. Rev. Lett.* **95**, 063601 (2005).
19. Alexander, A. L., Longdell, J. J., Sellars, M. J. & Manson, N. B. Photon echoes produced by switching electric fields. *Phys. Rev. Lett.* **96**, 043602 (2006).
20. Hétet, G., Longdell, J. J., Alexander, A. L., Lam, P. K. & Sellars, M. J. Electro-optic quantum memory for light using two-level atoms. *Phys. Rev. Lett.* **100**, 023601 (2008).
21. Staudt, M. U. *et al.* Fidelity of an optical memory based on stimulated photon echoes. *Phys. Rev. Lett.* **98**, 113601 (2007).
22. Staudt, M. U. *et al.* Interference of multimode photon echoes generated in spatially separated solid-state atomic ensembles. *Phys. Rev. Lett.* **99**, 173602 (2007).

23. Ohlsson, N., Nilsson, M. & Kröll, S. Experimental investigation of delayed self-interference for single photons. *Phys. Rev. A* **68,** 063812 (2003).
24. Hasegawa, Y. *et al.* Phase transfer in time-delayed interferometry with nuclear resonant scattering. *Phys. Rev. Lett.* **75,** 2216–2219 (1995).
25. Mitsunaga, M., Yano, R. & Uesugi, N. Spectrally programmed stimulated photon echo. *Opt. Lett.* **16,** 264–266 (1991).
26. Hastings-Simon, S. R. *et al.* Spectral hole burning spectroscopy in Nd^{3+}:YVO_4. *Phys. Rev. B* **77,** 125111 (2008).
27. Hesselink, W. H. & Wiersma, D. A. Picosecond photon echoes stimulated from an accumulated grating. *Phys. Rev. Lett.* **43,** 1991–1994 (1979).
28. Rippe, L., Julsgaard, B., Walther, A., Ying, Y. & Kröll, S. Experimental quantum-state tomography of a solid-state qubit. *Phys. Rev. A* **77,** 022307 (2008).
29. de Seze, F., Lavielle, V., Lorgeré, I. & Le Gouët, J. L. Chirped pulse generation of a narrow absorption line in a Tm3+:YAG crystal. *Opt. Commun.* **223,** 321–330 (2003).
30. Gorshkov, A. V., André, A., Fleischhauer, M., Sørensen, A. S. & Lukin, M. D. Universal approach to optimal photon storage in atomic media. *Phys. Rev. Lett.* **98,** 123601 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Cavalli and M. Bettinelli for kindly lending us the Nd:YVO_4 crystal. This work was supported by the Swiss NCCR Quantum Photonics and by the European Commission under the Integrated Project Qubit Applications.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.d.R. (hugues.deriedmatten@unige.ch).

Vol 456 | 11 December 2008 | doi:10.1038/nature07592

LETTERS

Enantiodivergent conversion of chiral secondary alcohols into tertiary alcohols

Jake L. Stymiest[1], Viktor Bagutski[1], Rosalind M. French[1] & Varinder K. Aggarwal[1]

From receptors in the nose to supramolecular biopolymers, nature shows a remarkable degree of specificity in the recognition of chiral molecules, resulting in the mirror image arrangements of the two forms eliciting quite different biological responses[1,2]. It is thus critically important that during a chemical synthesis of chiral molecules only one of the two three-dimensional arrangements is created. Although certain classes of chiral molecules (for example secondary alcohols[3,4]) are now easy to make selectively in the single mirror image form, one class—those containing quaternary stereogenic centres (a carbon atom with four different non-hydrogen substituents)—remains a great challenge[5–8]. Here we present a general solution to this problem which takes easily obtainable secondary alcohols in their single mirror image form and in a two-step sequence converts them into tertiary alcohols (quaternary stereogenic centres). The overall process involves removing the hydrogen atom (attached to carbon) of the secondary alcohol and effectively replacing it with an alkyl, alkenyl or aryl group. Furthermore, starting from a single mirror image form of the secondary alcohol, either mirror image form of the tertiary alcohol can be made with high levels of stereocontrol. Thus, a broad range of tertiary alcohols can now be easily made by this method with very high levels of selectivity. We expect that this methodology could find widespread application, as the intermediate tertiary boronic esters can potentially be converted into a range of functional groups with retention of configuration.

We live in a chiral world. Indeed, chirality and life are so inextricably linked that the detection of chirality outside our planet is used as a test for extraterrestrial life itself[9,10]. Nature's inherent chirality results in extraordinary specificity in the recognition of chiral molecules. For example, the different smell of oranges and lemons comes from two molecules identical except for their three-dimensional spatial arrangement. The classic example of the drug 'thalidomide' further illustrates the difference in biological response: one of the two enantiomers caused fetal birth defects (it restricted the development of blood vessels in rapidly dividing tissues and as such is now being used against certain forms of cancer[11]) whereas the mirror image form had the desired sedative properties[12]. However, in this case the tragedy could not have been avoided by administering the enantiomer with the desired properties, as the drug readily racemizes *in vivo*.

Because of the established importance of the three-dimensional spatial relationship between a small molecule (for example a drug) and its site of action (for example a protein), it has become necessary to test individual enantiomers and not mixtures of the two forms. A mixture of enantiomers can be difficult to separate, so it is important to control which mirror image form of a chiral molecule is created during its synthesis. This field, known as asymmetric synthesis, has grown over the past half century to a point where the synthesis of certain classes of compounds (for example diols[13] and secondary alcohols[3,4]) is now facile. However, chiral molecules containing quaternary stereogenic centres (for example tertiary alcohols)[5–8] remain difficult to make with high enantioselectivity.

Tertiary alcohols are commonly obtained by addition of an organometallic reagent to a ketone (Fig. 1, paths a and b). Although such a process can be rendered asymmetric through the use of chiral ligands[14–16], the process does not often lend itself to high asymmetric induction as it invariably relies on the steric difference between the substituents flanking the carbonyl group. When this difference is small, as in the case of ketones, the ratio of enantiomers formed is often low.

We therefore considered a very different concept to prepare quaternary stereogenic centres that makes use of the 1,2-metallate rearrangement of boronate complexes. We recognized that reaction of a chiral carbenoid bearing an alkyl and an aryl group should react stereoselectively with a boron reagent to give a chiral 'ate' complex. This species would subsequently undergo a stereospecific 1,2-metallate rearrangement to give a homologated boron intermediate bearing a quaternary stereogenic centre. Oxidation would then furnish a tertiary alcohol. As the process no longer relies on the different steric properties of similar-sized groups, but on whether the reaction of the chiral carbenoid occurs with retention or inversion[17,18], the potential for high enantioselectivity is greatly increased (Fig. 2).

A similar strategy has been previously attempted through reactions of chiral α-chloroboronic esters with Grignard reagents and α-chlorolithium reagents with chiral boronic esters[19]. Although such reaction partners can be successfully combined with high enantioselectivity in the synthesis of secondary alcohols[20], generation of sterically more demanding tertiary alcohols was much less rewarding: variable levels of selectivity were obtained and even the sense of asymmetric induction was unpredictable[19]. Clearly, the choice of the chiral carbenoid is critical. Our earlier studies had shown that Hoppe's lithiated carbamates[21] derived from primary alcohols reacted efficiently with boranes and boronic esters furnishing secondary alcohols and amines with high enantiomeric ratios[22]. We therefore considered the possibility of extending this process to the use of chiral carbenoids derived from secondary alcohols (many are commercially available in enantiomerically enriched form but otherwise are easily made by Noyori transfer hydrogenation[23]) to prepare tertiary alcohols.

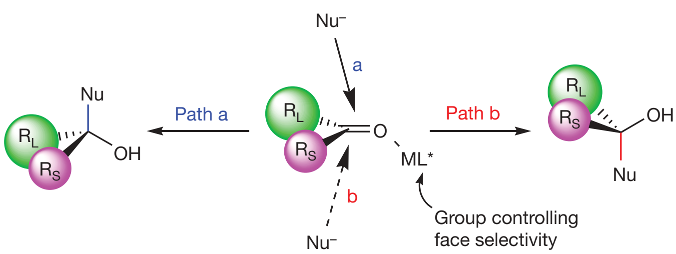


Figure 1 | Common strategy for preparing chiral tertiary alcohols through face selective addition of nucleophiles to ketones. Nu^-, nucleophile.

[1]School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK.

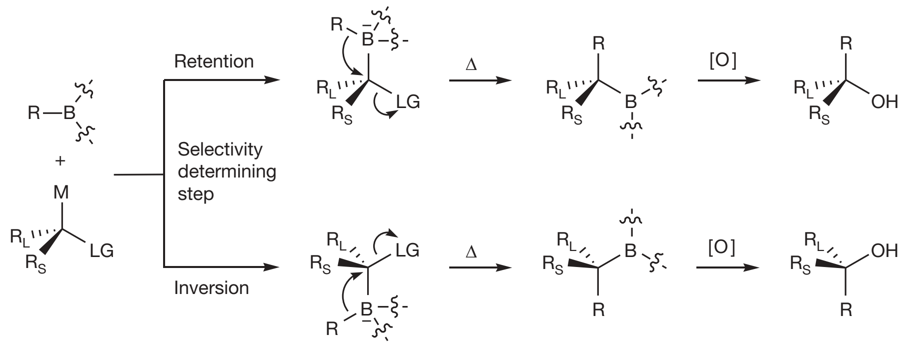
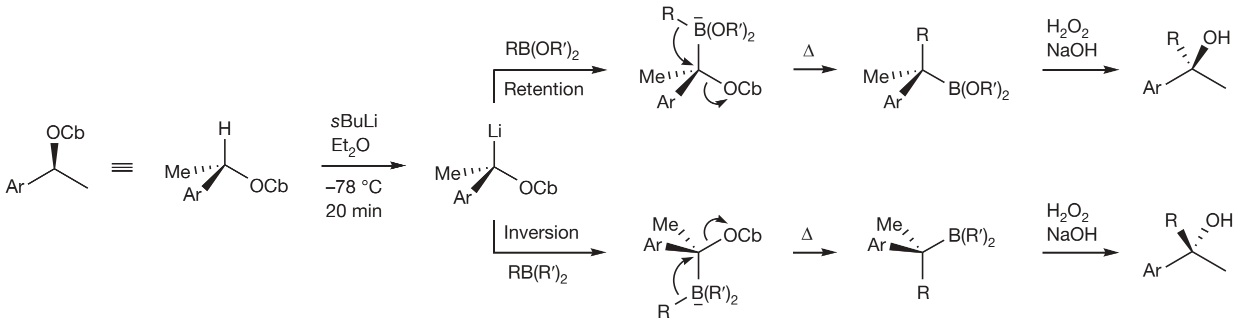


Figure 2 | Proposed synthesis of tertiary alcohols through addition of chiral carbenoids to boron reagents. Selectivity determined by degree of retention versus inversion in addition step. LG, leaving group; M = Li or Mg.

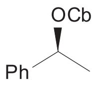
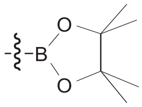
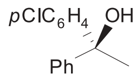
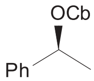
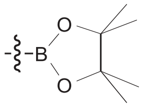
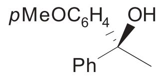
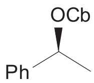
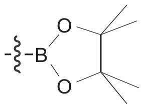
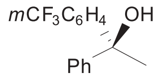
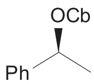
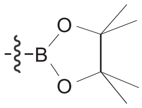
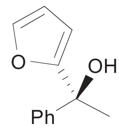
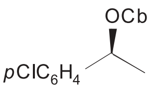
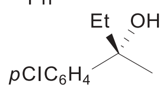
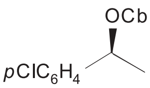
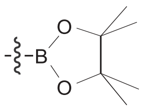
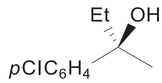
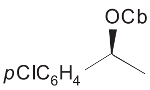
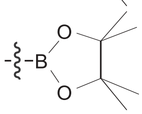
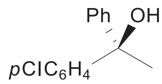
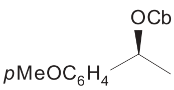
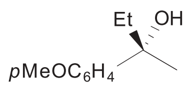
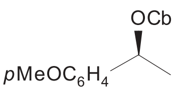
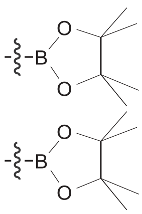
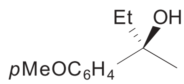
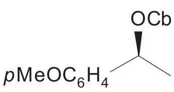
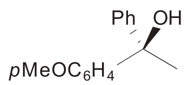
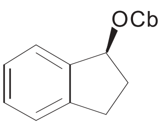
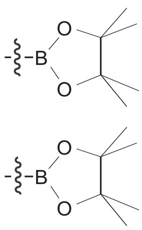
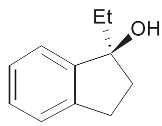
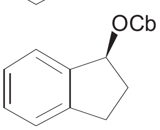
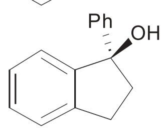
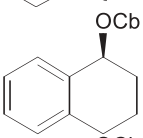
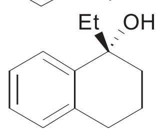
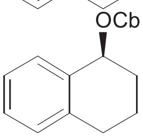
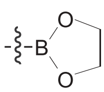
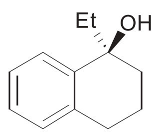
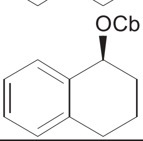
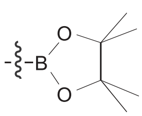
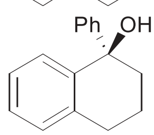
To test the concept, commercially available (*S*)-1-phenylethanol was converted into the corresponding *N*,*N*-diisopropyl carbamate. This carbamate had previously been deprotonated with *s*BuLi-TMEDA complex and reacted with a range of electrophiles[17]. During our investigation we found that *s*BuLi in Et_2O was sufficient on its own to effect deprotonation of the carbamate. Subsequent addition of the borane or boronic ester followed by oxidation gave the tertiary alcohol in good yield and very high enantiomeric ratio (Table 1). The enantiomeric

Table 1 | Lithiation–borylation of chiral secondary carbamates leading to tertiary alcohols



Entry	Carbamate (e.r.)	Migrating group, R	Borane/boronic ester component	Product	Yield (%) (e.r., S:R)
1	(99:1)	Et	BEt_2	Et, OH, Ph	91 (99:1)
2	(99:1)	Et	B (pinacol ester)	Et, OH, Ph	95 (1:99)
3	(99:1)	*i*Pr	B (9-BBN)	OH, Ph	91 (98:2)*
4	(99:1)	*i*Pr	B (pinacol ester)	OH, Ph	80 (4:96)
5	(99:1)	*n*Hex	B (9-BBN)	*n*Hex, OH, Ph	60 (98:2)*
6	(99:1)	*n*Hex	B (pinacol ester)	*n*Hex, OH, Ph	85 (4:96)
7	(99:1)	*c*Pr	B (pinacol ester)	OH, Ph	85 (3:97)
8	(99:1)	vinyl	B (pinacol ester)	OH, Ph	75 (2:98)
9	(99:1)	allyl	B (pinacol ester)	OH, Ph	95 (1:99)†

Table 1 | (Continued)

Entry	Carbamate (e.r.)	Migrating group, R	Borane/boronic ester component	Product	Yield (%) (e.r., S:R)
10	 (99:1)	pCl-C$_6$H$_4$-			97 (99:1)‡
11	 (99:1)	pMeO-C$_6$H$_4$-			92 (98:2)‡
12	 (99:1)	mCF$_3$-C$_6$H$_4$-			92 (99:1)‡
13	 (99:1)	2-furyl			94 (98:2)‡
14	 (98:2)	Et	-BEt$_2$		82 (95:5)
15	 (98:2)	Et			92 (4:96)§
16	 (98:2)	Ph			89 (4:96)†
17		Et	-BEt$_2$		87 (96:4)‖
18		Et			97 (2:98)‖
19		Ph			81 (4:96)‡‖
20		Et			69 (99:1)§‖
21		Ph			73 (6:94)‡‖
22		Et	-BEt$_2$		90 (5:95)‖
23		Et			98 (91:9)§‖¶
24		Ph			97 (4:96)†‖

The reactions were carried out in Et$_2$O in the presence of 1.1 equivalents of sBuLi, 1.1 equivalents of borane or 1.5 equivalents of boronic ester at the initial substrate concentration of 0.25 mol l^{-1} (method A). The following abbreviations are used: Me, methyl; Et, ethyl; iPr, isopropyl; nHex, n-hexyl; Ph, phenyl; Cb, N,N-diisopropylcarbamoyl.

* Method B was used in the oxidative work-up.

† Two equivalents of boronic ester were used.

‡ 1.1 equivalents of boronic ester were used.

§ Three equivalents of boronic ester were used.

‖ Deprotonation was carried out over 5 min in the presence of 1.1 equivalents of TMEDA (method C).

¶ One equivalent of LaCl$_3$·2LiCl (ref. 30) was added at −78 °C before warming up to ambient temperature. Use of three equivalents of EtB(pin) gave ~80:20 enantiomeric ratio (e.r.) with or without LaCl$_3$·2LiCl.

ratio of the starting carbamate is given to show the maximum enantiomeric ratio of the tertiary alcohol that is attainable.

In the case of boranes, both symmetrical trialkyl boranes and 9-BBN derivatives were examined, and good yields and high enantiomeric ratios were observed in all cases (Table 1, entries 1, 3 and 5). Interestingly, when the 9-BBN derivatives were used (entries 3 and 5), the boron substituent, rather than the boracycle, migrated exclusively, as was observed in the case of the less hindered primary carbamates[24,25].

We explored a broader range of boronic esters, including alkyl (entries 2, 4 and 6), cyclopropyl (entry 7), vinyl (entry 8), allyl (entry 9), aryl (entries 10, 11 and 12) and heterocyclic substrates (entry 13), as many of these are commercially available. In all cases high levels of selectivity were observed. Interestingly, the 1,2-metallate rearrangement of the boronate complex (established for the example in entry 2) readily occurred at 0 °C whereas related reactions involving lithiated carbamates derived from primary alkyl alcohols required higher temperatures (35 °C) and an additional Lewis acid ($MgBr_2$) to trigger the rearrangement[22].

Perhaps the most intriguing outcome was that in all cases the reactions occurred with almost complete inversion of stereochemistry when boranes were used but almost complete retention of stereochemistry when boronic esters were used. For a broad set of reactions to occur with such complete and yet opposite selectivity for the two sets of related reagents is unique. The consequence is that from a single enantiomer of a secondary alcohol, either enantiomer of a tertiary alcohol can now be prepared in high yield with very high enantioselectivity.

The origin of this difference in selectivity between the two reagents can be rationalized as follows (Fig. 3)[17,18]. In the case of boronic esters, the oxygen of the ester complexes with the lithium of the metallated carbamate and so is delivered on the same face as the metal. In the absence of such complexation, as in the case of the boranes, reaction occurs on the face opposite to the metal where there is significant electron density owing to the partially flattened nature of the carbanion. However, it should be noted that reactions of lithiated alkylcarbamates derived from non-benzylic primary alcohols occur with complete retention of stereochemistry with both boranes and boronic esters, presumably because in this case the non-mesomerically stabilized carbanion is essentially sp^3 hybridized and has very little electron density opposite the metal.

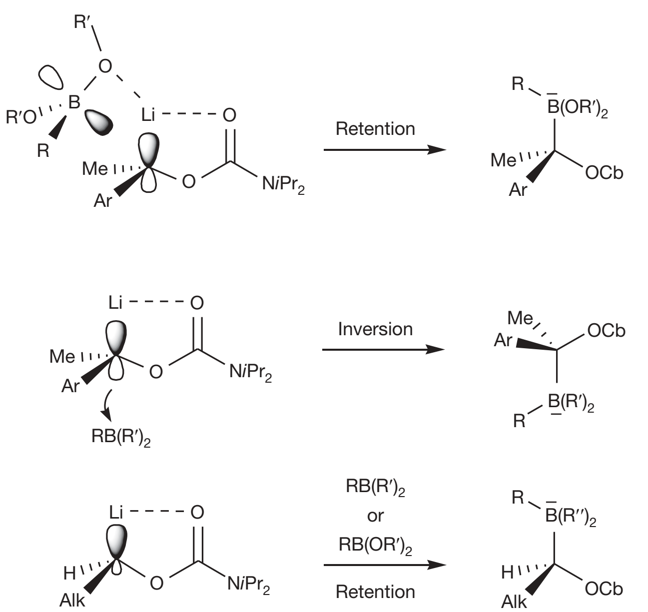


Figure 3 | Rationalization of the inversion versus retention of stereochemistry observed in reactions of lithiated carbamates with boranes and boronic esters respectively. Me, methyl; *i*Pr, isopropyl; Cb, *N*,*N*-diisopropylcarbamoyl.

In addition to the broad range of boron reagents tested, we examined alternative secondary alcohols bearing both electron-rich (pMeO-C_6H_4-) and electron-deficient (pCl-C_6H_4-) aromatics. These were prepared by Noyori asymmetric transfer hydrogenation[23] of the corresponding ketones and subsequently converted into the required carbamates. These substrates worked just as well with both boranes and boronic esters (entries 14–19), demonstrating the scope of this sequence. Once again, reactions with boranes occurred with inversion whereas boronic esters occurred with retention; both sets of reactions proceeding with high and opposite selectivity. The process was also applied to carbamates derived from the commercially available cyclic secondary alcohols, (*S*)-1-indanol and (*S*)-1-tetralol with a range of boranes and boronic esters (entries 20–24), thus extending the scope of the methodology to this additional class of substrates.

Two points are worth noting. First, aryl boranes, such as Ph-9-BBN, could not be used with these carbamates because protodeboronation of the resulting homologated boranes occurred extensively during the aqueous oxidative work-up. In contrast, no protodeboronation occurred during oxidative work-up of aryl boronic ester-derived homologated products, thus leading to diarylsubstituted tertiary alcohols in good yields (entries 10–13, 16, 19, 21 and 24). Second, in the case of the indanol-derived carbamate, reaction with triethylborane or the corresponding boronic ester lead to the same enantiomer. This was not unexpected as this substrate had been shown to react consistently with retention of stereochemistry with a range of electrophiles, as a result of increased pyramidalization of the geometrically constrained carbanion[26] (as discussed above). In any case, opposite enantiomers of the indanyl-based tertiary alcohols are easily obtained from (*R*)-1-indanol which is also commercially available.

Although some of the tertiary alcohols presented in Table 1 can also be made with high enantioselectivity by titanium-catalysed addition of diorganozinc reagents to ketones[27–29], its scope is not as broad as that presented here. Tertiary alcohols with branched alkyl, vinyl and heteroaryl groups (entries 3–8 and 13), which are easily accessible in high enantiomeric ratio by our method, cannot be easily made by the alternative route. Furthermore, although a few diorganozinc reagents are commercially available, the vast number of commercially available boronic esters greatly aids the application of the current methodology.

The simple methodology that we have described converts the simplest and most readily accessible chiral functional group (secondary alcohols) into the most difficult to obtain chiral entities—quaternary stereogenic centres (tertiary alcohols)—with very high enantioselectivity. The process shows very broad substrate scope in terms of both the secondary alcohols and the boranes or boronic esters used, thus allowing access to a very broad range of tertiary alcohols. Furthermore, either enantiomer of the tertiary alcohol can be obtained with high enantioselectivity from the same enantiomer of the secondary alcohol, simply depending on whether a borane or boronic ester is used. The plethora of functional groups into which the intermediate boranes/boronic esters can potentially be converted adds considerably to this new methodology.

METHODS SUMMARY

Enantioenriched secondary alcohols were obtained from commercial sources or prepared from the corresponding acetophenones by Noyori asymmetric transfer hydrogenation. These were converted into *N*,*N*-diisopropyl carbamates, deprotonated with *s*BuLi in Et_2O at –78 °C, and then reacted with a borane or boronic ester (–78 °C). On warming, the intermediate 'ate' complexes underwent 1,2-migration to give the corresponding homologated boranes/boronic esters, which were oxidized *in situ* with an excess of alkaline hydrogen peroxide to afford chiral tertiary alcohols in good yields and high enantioselectivities.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 March; accepted 27 October 2008.

1. Crossley, R. The relevance of chirality to the study of biological activity. *Tetrahedron* **48,** 8155–8178 (1992).
2. Thayer, A. M. Centering on chirality. *Chem. Eng. News* **85,** 11–19 (2007).

3. Pu, L. & Yu, H.-B. Catalytic asymmetric organozinc additions to carbonyl compounds. *Chem. Rev.* **101**, 757–824 (2001).
4. Noyori, R. & Hashiguchi, S. Asymmetric transfer hydrogenation catalysed by chiral ruthenium complexes. *Acc. Chem. Res.* **30**, 97–102 (1997).
5. Riant, O. & Hannedouche, J. Asymmetric catalysis for the construction of quaternary carbon centres: nucleophilic addition on ketones and ketimines. *Org. Biomol. Chem.* **5**, 873–888 (2007).
6. Christoffers, J. & Baro, A. (eds) *Quaternary Stereocentres: Challenges and Solutions for Organic Synthesis* (Wiley-VCH, 2005).
7. Corey, E. J. & Guzman-Perez, A. The catalytic enantioselective construction of molecules with quaternary carbon stereocentres. *Angew. Chem. Int. Ed.* **37**, 388–401 (1998).
8. Shibasaki, M. & Kanai, M. Asymmetric synthesis of tertiary alcohols and α-tertiary amines via Cu-catalysed C–C bond formation to ketones and ketimines. *Chem. Rev.* **108**, 2853–2873 (2008).
9. Thiemann, W. H.-P. Homochirality of the evolution of biospheres. *Jpn Soc. Biol. Sci. Space* **12**, 73–77 (1998).
10. Levin, G. B., Miller, J. D., Straat, P. A., Lodder, R. A. & Hoover, R. B. Detecting life and biology-related parameters on Mars. ⟨http://mars.spherix.com/IEEE%20Pape-FINAL.pdf⟩ (2006).
11. D'Amato, R. J., Loughnan, M. S., Flynn, E. & Folkman, J. Thalidomide is an inhibitor of angiogenesis. *Proc. Natl Acad. Sci. USA* **91**, 4082–4085 (1994).
12. Rouhi, M. Top pharmaceuticals: thalidomide. *Chem. Eng. News* **83**, 122–123 (2005).
13. Kolb, H. C., VanNieuwenhze, M. S. & Sharpless, K. B. Catalytic asymmetric dihydroxylation. *Chem. Rev.* **94**, 2483–2547 (1994).
14. Chen, C.-A., Wu, K.-H. & Gau, H.-M. Highly enantioselective aryl additions of [$AlAr_3$(thf)] to ketones catalyzed by a titanium(IV) catalyst of (*S*)-binol. *Angew. Chem. Int. Ed.* **46**, 5373–5376 (2007).
15. Ramón, D. J. & Yus, M. First enantioselective addition of dialkylzinc to ketones promoted by titanium(IV) derivatives. *Tetrahedr. Lett.* **39**, 1239–1242 (1998).
16. Betancort, J. M., García, C. & Walsh, P. J. Development of the first practical catalyst for the asymmetric addition of alkyl- and arylzinc reagents to ketones. *Synlett* **2004**, 749–760 (2004).
17. Hoppe, D., Carstens, A. & Kramer, T. Generation of a configurationally stable chiral benzyllithium derivative, and the capricious stereochemistry of its electrophilic substitution. *Angew. Chem. Int. Edn Engl.* **29**, 1424–1425 (1990).
18. Carstens, A. & Hoppe, D. Generation of a configurationally stable, enantioenriched α-oxy-α-methylbenzyllithium: stereodivergence of its electrophilic substitution. *Tetrahedron* **50**, 6097–6108 (1994).
19. Matteson, D. S. & Hurst, G. D. Asymmetric synthesis of tertiary alcohols from α-halo boronic esters. *Heteroatom Chem.* **1**, 65–74 (1990).
20. Matteson, D. S. α-Halo boronic esters in asymmetric synthesis. *Tetrahedron* **54**, 10555–10607 (1998).
21. Hoppe, D. & Hense, T. Enantioselective synthesis with lithium/(−)-sparteine carbanion pairs. *Angew. Chem. Int. Edn Engl.* **36**, 2282–2316 (1997).
22. Stymiest, J. L., Dutheuil, G., Mahmood, A. & Aggarwal, V. K. Lithiated carbamates: chiral carbenoids for iterative homologation of boranes and boronic esters. *Angew. Chem. Int. Ed.* **46**, 7491–7494 (2007).
23. Fujii, A., Hashiguchi, S., Uematsu, N., Ikariya, T. & Noyori, R. Ruthenium(II)-catalyzed asymmetric transfer hydrogenation of ketones using a formic acid-triethylamine mixture. *J. Am. Chem. Soc.* **118**, 2521–2522 (1996).
24. Aggarwal, V. K., Fang, G. Y., Ginesta, X., Howells, D. & Zaja, M. Toward an understanding of the factors responsible for the 1,2-migration of alkyl groups in borate-complexes. *Pure Appl. Chem.* **78**, 215–229 (2006).
25. Fang, G. Y. *et al.* Asymmetric sulfur ylide reactions with boranes: scope and limitations, mechanism and understanding. *J. Am. Chem. Soc.* **129**, 14632–14639 (2007).
26. Derwing, C., Frank, H. & Hoppe, D. Chiral benzyllithium compounds: high configurative stability of (*R*)- and (*S*)-1-lithioindan-1-yl *N,N*-diisopropylcarbamate and unexpected stereochemical course of the substitution reactions. *Eur. J. Org. Chem.* **1999**, 3519–3524 (1999).
27. Jeon, S.-J., Li, H., García, C., LaRochelle, L. K. & Walsh, P. Catalytic asymmetric addition of alkylzinc and functionalized alkylzinc reagents to ketones. *J. Org. Chem.* **70**, 448–455 (2005).
28. Jeon, S.-J., Ji, H. & Walsh, P. J. A green chemistry approach to a more efficient asymmetric catalyst: solvent-free and highly concentrated alkyl additions to ketones. *J. Am. Chem. Soc.* **127**, 16416–16425 (2005).
29. Forrat, V. J., Ramón, D. J. & Yus, M. Chiral tertiary alcohols from a *trans*-1-arenesulfonyl-amino-2-isoborneolsulfonylaminocyclohexane-catalyzed addition of organozincs to ketones. *Tetrahed. Asymm.* **16**, 3341–3344 (2005).
30. Krasovskiy, A., Kopp, F. & Knochel, P. Soluble lanthanide salts ($LnCl_3 \cdot 2\,LiCl$) for the improved addition of organomagnesium reagents to carbonyl compounds. *Angew. Chem. Int. Ed.* **45**, 497–500 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank EPSRC for financial support and Frontier Scientific for their donation of boronic esters. We also thank Merck for unrestricted research support. V.K.A. thanks the Royal Society for a Wolfson Research Merit Award and EPSRC for a Senior Research Fellowship.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to V.K.A. (v.aggarwal@bristol.ac.uk).

METHODS

Method A. To a stirred solution of the respective carbamate (1.00 mmol) in anhydrous Et_2O (4.0 ml) at –78 °C (dry ice/acetone bath) under argon, *s*BuLi (0.82 ml of 1.3 M solution, 1.06 mmol) was added dropwise over 2 min. This mixture was allowed to stir at –78 °C for 20 min, and then 1 M solution of the respective borane (1.06 ml, 1.06 mmol) or boronic ester (1.1–3 ml, 1.1–3.0 mmol, see Table 1) in diethyl ether was added dropwise over 2 min with vigorous stirring. The reaction mixture was kept at –78 °C for an additional 30 min, then the cooling bath was removed and the reaction mixture was stirred at ambient temperature for 2 h. The reaction mixture was diluted with anhydrous THF (4 ml) containing ~5 mg of 2,6-di-*tert*-butyl-4-methylphenol, cooled to ~2 °C (water-ice bath), and an ice-cold degassed mixture of 3 M NaOH (1.4 ml) and 30% aqueous H_2O_2 (0.8 ml) per 1 mmol of borane or boronic ester used was added all at once. The cooling bath was removed and the reaction mixture was stirred at ambient temperature for 2 h, then diluted with water (5 ml) and extracted with diethyl ether (3 × 10 ml). The combined organic phases were washed with brine (30 ml), dried with $MgSO_4$ and concentrated *in vacuo*. The crude product was purified by flash chromatography (PE/EtOAc) to afford pure tertiary alcohol.

Method B. Method A was followed, except that after the $NaOH/H_2O_2$ mixture was added, the reaction flask was fitted with a reflux condensor and stirring was continued at 60 °C (bath temperature) for 1 h, after which 1 ml of 30% H_2O_2 was added, and the mixture was refluxed for an additional 2 h. The remaining work-up was as described above.

Method C. Method A was followed, except that the deprotonation of a carbamate was performed in the presence of 1.1 equiv. of TMEDA (165 µl, 1.1 mmol) within 5 min.

LETTERS

Strain accommodation by slow slip and dyking in a youthful continental rift, East Africa

Eric Calais[1], Nicolas d'Oreye[2], Julie Albaric[3], Anne Deschamps[4], Damien Delvaux[5], Jacques Déverchère[3], Cynthia Ebinger[6], Richard W. Ferdinand[7], François Kervyn[5], Athanas S. Macheyeki[8,9], Anneleen Oyen[2,10], Julie Perrot[3], Elifuraha Saria[11], Benoît Smets[5], D. Sarah Stamps[1] & Christelle Wauthier[5,12]

Continental rifts begin and develop through repeated episodes of faulting and magmatism, but strain partitioning between faulting and magmatism during discrete rifting episodes remains poorly documented. In highly evolved rifts, tensile stresses from far-field plate motions accumulate over decades before being released during relatively short time intervals by faulting and magmatic intrusions[1–3]. These rifting crises are rarely observed in thick lithosphere during the initial stages of rifting. Here we show that most of the strain during the July–August 2007 seismic crisis in the weakly extended Natron rift, Tanzania, was released aseismically. Deformation was achieved by slow slip on a normal fault that promoted subsequent dyke intrusion by stress unclamping. This event provides compelling evidence for strain accommodation by magma intrusion, in addition to slip along normal faults, during the initial stages of continental rifting and before significant crustal thinning.

In July–August 2007, a seismo-magmatic crisis in the Natron basin (Fig. 1) was accompanied by the first dyking event ever captured geodetically in a continental rift[4]. The <5-Myr-old Natron basin lies near the southern termination of the eastern branch of the East African Rift, the divergent boundary between the Somalian and Nubian plates. Alkaline and basaltic volcanism was coeval with initial surface faulting[5], yet strain accommodation by magma intrusion is virtually unconstrained. With current far-field extension rates on the order of 3 to 4 mm yr^{-1} (ref. 6), the eastern rift has experienced only a few kilometres of crustal thinning[7], whereas lithospheric thickness is ~90 km relative to >125-km-thick lithosphere outside the rift[8,9]. The Natron crisis therefore provides a rare opportunity to quantify how faulting and magma intrusion contribute to strain accommodation within a youthful continental rift.

The Natron seismo-magmatic crisis began on 12 July 2007 with two earthquakes of respective body-wave magnitudes $M_b = 4.4$ and $M_b = 4.5$ recorded at the southern end of Lake Natron (Fig. 1). Subsequent seismic activity included an $M_b = 5.3$ event and an $M_b = 4.6$ event and culminated on 17 July with a shallow (4.7-km depth) event with moment magnitude $M_w = 5.9$. Seismic activity continued in July (32 $M_b > 3.8$ events) and August (21 $M_b > 3.9$ events) and then decayed rapidly in September (four $M_b > 4.2$ events). Owing to the fortuitous deployment of a temporary broadband seismic network just to the south of the Natron basin, we were able to locate the 17 July earthquake and 607 events for the interval 12–31 July located ~5–10 km south of the Gelai shield volcano (Fig. 1). Seismicity started on 12 July below the southern toe of Gelai at depths between 4 and 12 km, and then migrated up to 10 km to the northeast

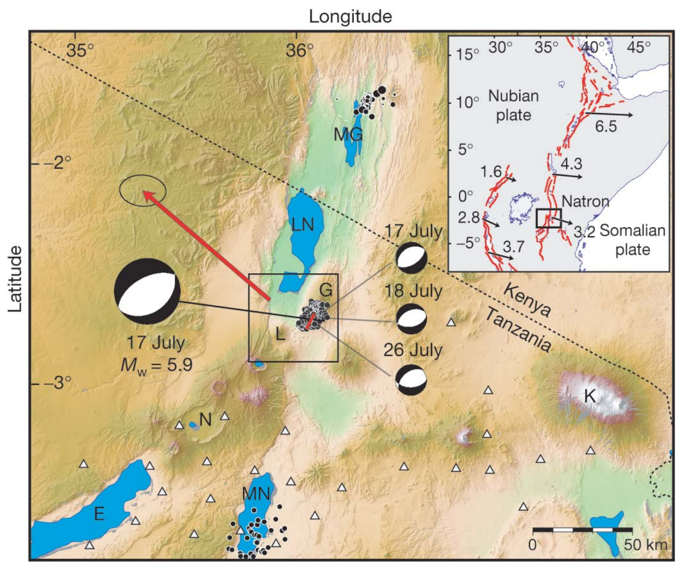


Figure 1 | Regional setting. Black circles show epicentres of earthquakes with body-wave magnitudes $M_b > 3.8$ for the 16 July–21 August 2007 interval. Black and white circles indicate focal mechanisms for the largest magnitude events in that time interval. Black circles near lakes Magadi (MG) and Manyara (MN) show the 1998 and 1994–1995 seismic swarms[17,19]. The short red line shows the location of the dyke intruded along the southern flank of Gelai volcano (G). White triangles show the seismic stations of the SEISMO-TANZ'07 temporary deployment. The background colour shows topography based on data from NASA's Shuttle Radar Topography Mission. The red arrow indicates GPS displacement at a geodetic benchmark at the southern end of Lake Natron (LN), installed and observed for the first time in August 2006 and re-observed during 4–7 August 2007. Processing of the dual-frequency phase data shows a displacement of 5.6 ± 0.9 cm vertically and 5.7 ± 0.4 cm horizontally in the N49 W direction. Focal mechanisms for the 17 July event and earthquake locations are from this study (see Methods). Focal mechanisms for smaller magnitude events are from the Harvard Centroid-Moment Tensor catalogue. Black rectangle (in main panel) outlines the area of Fig. 3. E, Lake Eyasi; L, Ol Doinyo Lengai; K, Mount Kilimanjaro; N, Ngorongoro crater. The inset shows the study area (black rectangle) along the East African Rift. Arrows show predicted extension rates (in millimetres per year) and directions across the main rift structures[6].

[1]Purdue University, Department of Earth and Atmospheric Sciences, West Lafayette, Indiana 47906, USA. [2]National Museum of Natural History, Department of Geophysics/Astrophysics, L-2160 Luxembourg. [3]Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale, Plouzané 29280, France. [4]Géosciences Azur, UNS/CNRS, Valbonne 06560, France. [5]Royal Museum for Central Africa, Tervuren 3080, Belgium. [6]University of Rochester, Department of Earth and Environmental Sciences, New York, New York 14627, USA. [7]University of Dar es Salaam, Department of Geology, PO Box 35091, Dar es Salaam, Tanzania. [8]Renard Centre of Marine Geology, Department of Geology and Soil Science, Universteit Gent, Gent 9000, Belgium. [9]Mineral Resources (Madini) Institute, PO Box 903, Dodoma, Tanzania. [10]Delft University of Technology, Delft 2600, The Netherlands. [11]Ardhi University, Department of Geomatics, PO Box 35176, Dar es Salaam, Tanzania. [12]University of Liège, Liège 4000, Belgium.

between 14 and 17 July. Hypocentres delineate a steep plane dipping to the northwest, and striking N33 E close to the geometry of the focal mechanism of the main event (Fig. 2). Following that event, seismicity remained stationary under the southern slope of Gelai, with a broader distribution of hypocentres at depths between 1 and 14 km, but still clustered in the N33 E direction.

Field observations in early August and October showed metres- to kilometres-long, newly formed *en échelon* tension fractures along the southern flank of Gelai, delineating two main north-northeast–south-southwest strands separated by 2–3 km. Vertical displacements are consistently east-down for the western strand and west-down for the eastern strand, resulting in a narrow central graben. Displacements are larger along the eastern strand in its middle segment, with an average of ~12 cm of opening and ~35 cm of vertical offset. This strand also shows evidence for repeated activation in recent times. No eruption or venting was recorded at Gelai, but the crisis was accompanied by renewed activity at the nearby Ol Doinyo Lengai carbonatitic volcano[10] (Fig. 1), with explosive eruptions bearing similarities to the 1966 events[11].

Surface deformation associated with the Natron crisis was also quantified using data from the Global Positioning System (GPS; Fig. 1) and radar interferometry (InSAR; Fig. 3). An interferogram spanning 3 April to 17 July shows that deformation started before the main $M_w = 5.9$ event, with ellipse-shaped fringes that coincide with earthquake epicentres for that time interval (Fig. 3). Their pattern, as well as the distribution of earthquake hypocentres and the focal mechanism of the larger events, is consistent with slip on a buried, northwest-dipping normal fault. The 17 July–21 August interferogram (Fig. 3c) shows deformation-related fringes centred on the southern flank of Gelai in a pattern indicative of the subsidence of a fault-bounded central graben bounded on either side by areas of uplift and horizontal displacement away from the graben. This is consistent with the deformation expected from a blade-shaped dyke with induced surface faulting at its upper tip[3,12]. At least four fringes (~12 cm in range) are also visible on the eastern flank of Gelai, indicative of the subsidence possibly induced by magma chamber deflation.

We model surface deformation as the result of dislocations on faults[13] and volume changes at point sources (so-called 'Mogi sources'[14]) buried in an elastic half-space (see Methods) to estimate the amount of slip and opening on the faults involved and the volume changes due to the dyke intrusion and magma chamber deflation. We first model the 3 April–17 July interferogram, using a normal fault with strike and dip taken from the geometry of the plane derived from the 12–17 July hypocentres (Fig. 3b). The best-fit solution shows up to 0.5 m of normal slip concentrated at a depth of between 5 and 10 km, with a distribution that matches earthquake hypocentres, and a total moment release of 2.7×10^{18} N m. Because the cumulative seismic moment release for that time interval is 6.5×10^{16} N m, ~98% of the total moment must have been dissipated aseismically. Slip most likely occurred from 12 to 17 July, as seismicity was migrating from south to north along the fault (Fig. 2), in the first slow slip event recorded on a normal fault in a continental rift.

We then model the 17 July–21 August interferogram with a combination of fault slip and dyke opening. We use discontinuities in the interferograms and field observations to map the surface trace of the graben-bounding faults, where we allow slip down to depths of 2–4 km (from south to north, as the graben widens), and include the 'pre-17 July' normal fault defined above. We simulate the dyke as a vertical tensile dislocation with its surface projection following the centre of the axial graben and allow opening between the surface

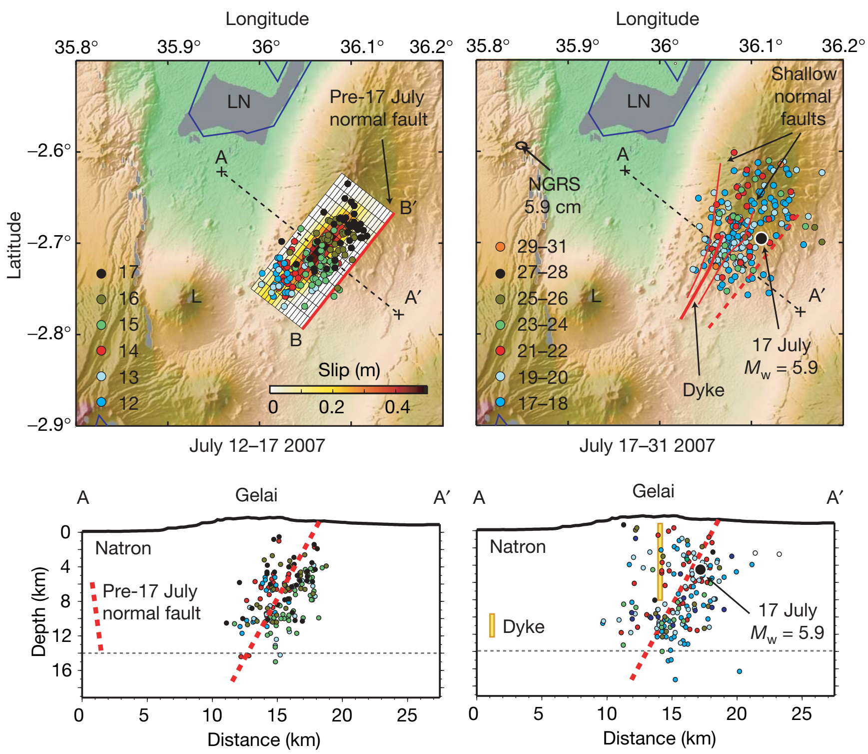


Figure 2 | Earthquake locations from the local seismic network. Earthquakes are colour-coded as a function of date; the key shows the corresponding days of July 2007. Top panels: map view; bottom panels: cross-sectional view. The thin dashed lines at 14-km depth indicate a velocity increase from 5.9 to 6.5 km s^{-1} for seismic P waves in the crustal model used for earthquake location (see Methods). The left-hand panels show the 12–17 July interval. Event depths increase from southeast to northwest and delineate a fault plane deepening to the northwest at 60° and striking N33 E (from a least-squares fit through hypocentres). Epicentres (top) are superimposed on the surface projection of the slip distribution estimated on this fault, showing that they coincide with the area of maximum slip. The right-hand panels show the interval from 17 July (the time of the $M_w = 5.9$ event, marked as a white-circled black dot) to 31 July. Dashed lines between points A and A′ indicate the cross-sections used in the bottom panels. NGRS indicates the location of the Engaresero GPS site, where a horizontal displacement of 5.7 cm was measured in the direction indicated by the arrow.

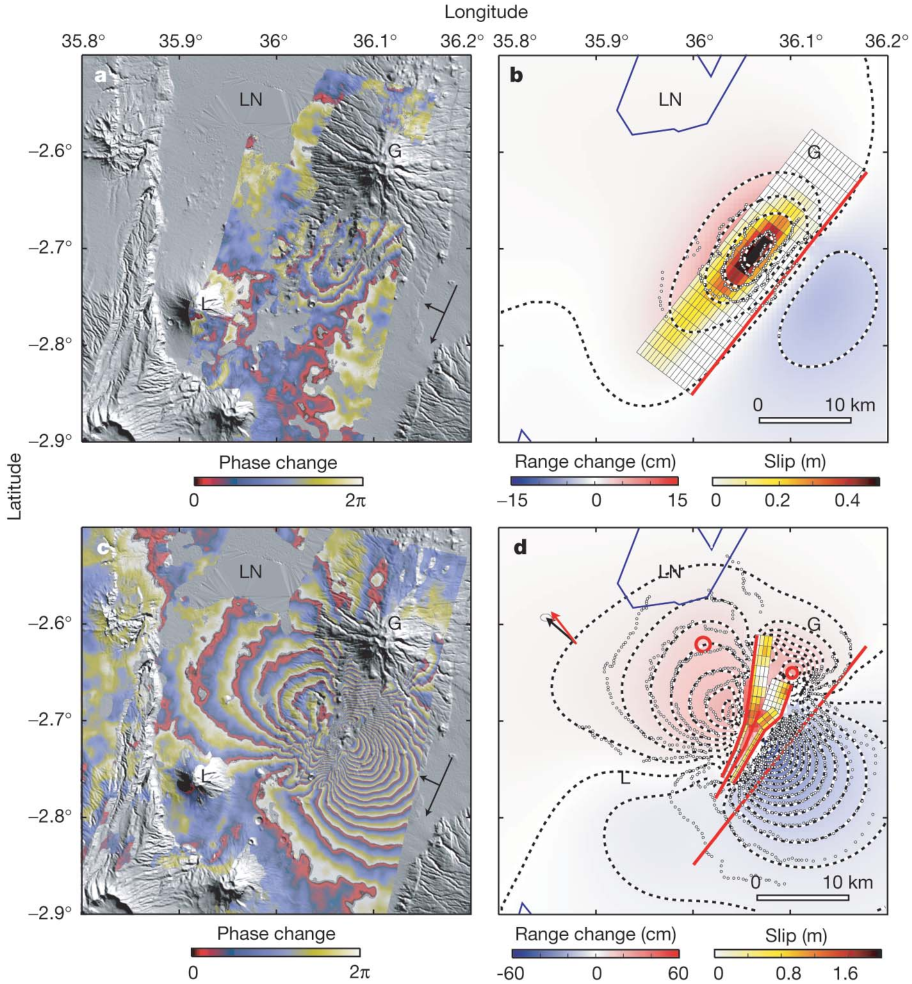


Figure 3 | Interferograms and corresponding model results. Each fringe in the interferograms (that is, phase change in cycles, reckoned positive for increasing range) represents a ground displacement of 2.8 cm in the satellite–ground direction. **a**, 7 April–17 July interferogram (the 17 July image was made 7 h before the $M_w = 5.9$ earthquake) showing four elliptical fringes indicating ~12 cm of positive range change (ground moving away from the satellite) and one poorly defined negative range change fringe (to the east). **b**, Corresponding model. **c**, 17 July–21 August interferogram showing deformation-related fringes covering an area of about $50 \times 70\,km^2$ centred on the southern flank of Gelai. The two-lobe fringe pattern indicates up to 45 cm of negative range change (ground moving towards the satellite) in the eastern half of the scene and 17 cm of positive range change in the western half. Fringes in the lobes converge along a 15-km-long, 2–3-km-wide elongated area of positive range change (≤55 cm in its southern part and >45 cm in its northern part). Fringes in this central area are not contiguous with the side lobes, but have discontinuities matching the fractures mapped in the field. **d**, Corresponding model. Estimated slip on the pre-17 July normal fault, shown in Fig. 4d, is omitted here for sake of clarity. In the interferograms, black orthogonal arrows show the satellite displacement and look directions. In the models, red lines show the intersections of model faults and dyke with the surface; grey dots delineate observed fringes; dashed lines show model fringes; background colours show range change; the thin grid shows the surface projection of the fault model; and colours in the grid show the distribution of slip and opening estimated in the inversion. Black and red arrows show observed and predicted displacement, respectively, at GPS site Engaresero. Red circles show the location of the Mogi sources under Gelai and the Natron basin. See Supplementary Figure 1 for model residuals.

and a depth of 12 km. We include volume-change point sources below Gelai to simulate the local subsidence in Fig. 3c and below the Natron basin to explain the asymmetry of the fringe lobe northwest of the dyke.

The best-fit inversion (Figs 3d, 4) includes ≤2 m of dyke opening between depths of ~2 and 6 km, centred below the southern half of the fault-bounded graben, and ≤0.6 m of slip on the shallow normal faults on either side of the dyke, consistent with field observations. We find a patch of shallow slip (4–5-km depth) on the pre-17 July fault, with a corresponding moment release of 8.5×10^{17} N m, or $M_w = 5.95$, consistent with the 17 July earthquake. The computation of stress changes caused by the pre-17 July slip acting on vertical faults trending N25 E shows that the area of largest estimated dyke opening coincides with the most negative change in normal stress (Fig. 3c). This suggests that dyke opening was triggered by static stress changes associated with pre-17 July slip (see, for example, ref. 15). We find a total volume of $0.09\,km^3$ of magma intrusion in the dyke and a negative volume change of $0.06\,km^3$ at the Mogi point sources ($0.01\text{-}km^3$ deflation of a 3-km-deep point source below Gelai volcano and $0.05\text{-}km^3$ deflation of a 10-km-deep point source west of Gelai). Some or all of the magma intrusion may therefore have its source in the magma chambers below and west of Gelai (see Methods).

The total moment released by fault slip and dyke opening estimated from geodetic data for the 17 July–21 August interval is 4.0×10^{18} N m. Because cumulative seismic moment release for that interval is 1.4×10^{18} N m, about 65% of the total moment must have been released by aseismic processes. Our inversion gives a moment of 2.6×10^{18} N m for the (aseismic) dyke opening, or 65% of the total 'geodetic' moment release, consistent with the missing moment inferred from seismic processes alone. This moment release budget illustrates the importance of aseismic processes to the July 2007 Natron rifting events, a feature consistent with magma-assisted oceanic rifting episodes[2,3]. The 2007 Natron rifting episode, therefore, provides a snapshot of strain partitioning between faulting and

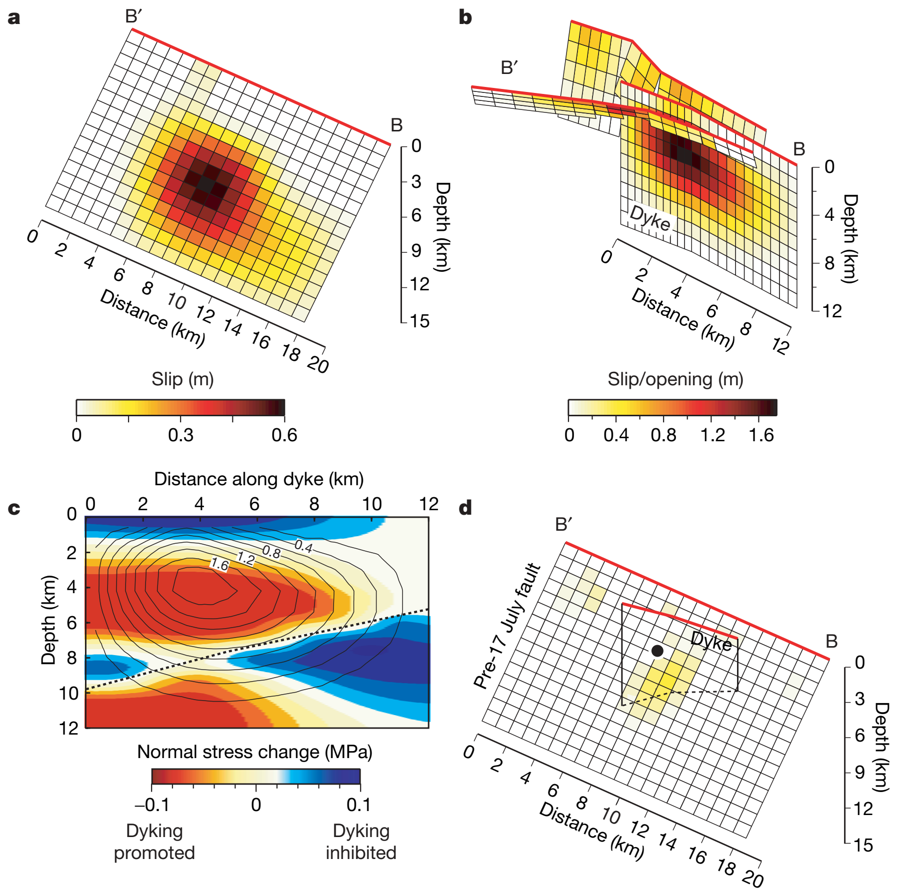


Figure 4 | Slip/opening model. a, Estimated slip distribution for the 12–17 July slow slip event on the northwest-dipping normal fault. The locations of B and B′ are shown in Fig. 2. **b**, Estimated opening distribution along the dyke and slip distribution on the shallow normal faults bounding the central graben. **c**, Normal stress changes (background colours) caused by pre-17 July slip, resolved along the (future) dyke. Negative normal stress changes (red) promote dyke opening (indicated in metres by the contours; see **b**). The black dashed line is the intersection of the pre-17 July normal fault with the dyke. **d**, Estimated slip distribution on the pre-17 July normal fault for the 17 July–21 August interval. The colour scale for the slip distribution is the same as in **b**. The shallow slip patch probably corresponds to the 17 July $M_w = 5.9$ earthquake, whose hypocentre is marked by a black circle. We show the geometry of the dyke, for reference. The dashed black line shows the intersection of the dyke with the pre-17 July fault plane. Red lines show the intersection of the modelled fault and dyke planes with the surface, shown in map view in Fig. 2.

magma intrusion processes needed to link discrete rifting episodes with the time-averaged deformation.

The largely aseismic Natron rifting event would have remained undetected in the absence of local seismic networks, surface volcanic eruptions or geodetically measured surface deformation. Indirect evidence, however, indicates that this event is typical of the rifting process in the eastern rift system, and that magma intrusions fundamentally modify the crust and upper-mantle structure during the initial stages of rifting. Magma intrusion is common in space and time, with volcanism preceding or concurrent with the onset of faulting in the eastern rift. In addition, although crustal thinning is less than 10%, mantle xenoliths show that the cratonic mantle lithosphere beneath the eastern rift has been metasomatized by magmatic fluids[16]. The earthquake sequence reported here bears similarities to other earthquake swarms in the eastern rift that occurred before there was adequate geodetic coverage[17–19]. Narrow, high-velocity zones in the upper crust are interpreted as cooled melt intrusion from earlier dyking events[17].

Magma-assisted rifting models provide a unified framework for these observations. Owing to the positive buoyancy of melt, dyke intrusions may reduce the force required to cause extensional yielding of initially thick continental lithosphere by a factor of up to six[20]. Dyke intrusions occur at lower tectonic stresses than faulting, provided that magma pressure is high[12]. The density contrast caused by the dyke intrusions, and the reduction in plate strength by magmatic heating over repeated rifting events, localizes strain to zones of previous dyke intrusions[20]. The combination of a slow slip event, an aseismic dyke intrusion and normal faulting in a young cratonic continental rift is direct evidence of the role of strain accommodation through magma intrusions during the initial stages of continental rifting.

METHODS SUMMARY

We used arrival times from a local seismic network to locate earthquakes in the Natron rift. Ground deformation was calculated using radar interferograms based on data from Envisat (an Earth-observation satellite of the European Space Agency (ESA)) and GPS measurements. Field geological mapping of the ground fracture system associated with the seismo-magmatic crisis complemented these geophysical measurements. Observational constraints from fieldwork and seismology were used to design the model geometry. Fault slip and dyke opening were then calculated from a kinematic inversion of the ground deformation data.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 July; accepted 25 September 2008.

1. Heki, K., Foulger, G., Julian, B. & Jahn, C. H. Plate dynamics near divergent boundaries: Geophysical implications of postrifting crustal deformation in NE Iceland. *J. Geophys. Res.* **98**, 14279–14297 (1993).
2. Cattin, R. *et al.* Numerical modelling of Quaternary deformation and post-rifting displacement in the Asal–Ghoubbet rift (Djibouti, Africa). *Earth Planet. Sci. Lett.* **239**, 352–367 (2005).
3. Wright, T. J. *et al.* Magma-maintained rift segmentation at continental rupture in the 2005 Afar dyking episode. *Nature* **442**, 291–294 (2006).

4. Baer, G., Hamiel, Y., Shamir, G. & Nof, R. Evolution of a magma-driven earthquake swarm and triggering of the nearby Oldoinyo Lengai eruption, as resolved by InSAR, ground observations, and elastic modelling, East African Rift, 2007. *Earth Planet. Sci. Lett.* **272,** doi:10.1016/j.epsl.2008.04.052 (2008).
5. Ebinger, C. J., Poudjom Djomani, Y., Mbede, E., Foster, A. N. & Dawson, J. B. Rifting Archean lithosphere: Eyasi–Manyara–Natron rifts, East Africa. *J. Geol. Soc. Lond.* **154,** 947–960 (1997).
6. Stamps, D. S. *et al.* A kinematic model for the East African Rift. *Geophys. Res. Letters* **35,** doi:10.1029/ 2007GL032781 (2008).
7. Birt, C. *et al.* The influence of pre-existing structures on the evolution of the southern Kenya rift valley – Evidence from seismic and gravity studies. *Tectonophysics* **278,** 211–242 (1997).
8. Green, W. V., Achauer, U. & Meyer, R. P. A three-dimensional seismic image of the crust and upper mantle beneath the Kenya rift. *Nature* **354,** 199–203 (1991).
9. Nyblade, A. A., Owens, T. J., Gurrola, H., Ritsema, J. & Langston, C. A. Seismic evidence for a deep upper mantle thermal anomaly beneath East Africa. *Geology* **7,** 599–602 (2000).
10. Vaughan, R. G., Kervyn, M., Realmuto, V., Abrams, M. & Hook, S. J. Satellite measurements of recent volcanic activity at Oldoinyo Lengai, Tanzania. *J. Volcanol. Geotherm. Res.* **173,** 196–206 (2008).
11. Dawson, J. B., Bowden, P. & Clark, G. C. Activity of the carbonatite volcano, Oldoinyo Lengai, 1966. *Geol. Rundsch.* **57,** 865–879 (1968).
12. Rubin, A. Dike-induced faulting and graben subsidence in volcanic rift zones. *J. Geophys. Res.* **92,** 1839–1858 (1992).
13. Okada, Y. Internal deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **82,** 1018–1040 (1992).
14. Mogi, K. Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. *Bull. Earthq. Res. Inst. Tokyo* **36,** 99–134 (1958).
15. Amelung, F., Yun, S.-H., Walter, T. R., Segall, P. & Kim, S. W. Stress control of deep rift intrusion at Mauna Loa volcano, Hawaii. *Science* **316,** 1026–1030 (2007).
16. Vauchez, A., Dineur, F. & Rudnick, R. Microstructure, texture and seismic anisotropy of the lithospheric mantle above a mantle plume: Insights from the Labait volcano xenoliths (Tanzania). *Earth Planet. Sci. Lett.* **232,** 295–314 (2005).
17. Ibs-von Seht, M., Blumenstein, S., Wagner, R., Hollnack, D. & Wohlenberg, J. Seismicity, seismotectonics and crustal structure of the southern Kenya Rift - new data from the Lake Magadi area. *Geophys. J. Int.* **146,** 439–453 (2001).
18. Tongue, J., Maguire, P. K. H. & Burton, P. An earthquake study in the Lake Baringo basin of the central Kenya Rift. *Tectonophysics* **236,** 151–164 (1994).
19. Nyblade, A. A., Birt, C., Langston, C. A., Owens, T. J. & Last, R. J. Seismic experiment reveals rifting of craton in Tanzania. *Eos* **77,** 517–521 (1996).
20. Buck, W. R. in *Rheology and Deformation of the Lithosphere at Continental Margins* (eds Karner, G. D., Taylor, B., Driscoll, N. & Kohlstedt, D.) 1–30 (Columbia Univ. Press, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Envisat/ASAR data were systematically programmed and acquired through the ESA Cat-1 project no. C1P.3224. Interferograms were computed using DORIS software (Delft University of Technology). Precise orbits were provided by the Delft Institute of Earth Observation and Space Systems and ESA. We thank the Tanzania Survey and Mapping Department for their support of the October field expedition and the Tanzania Commission for Science and Technology. We thank V. Cayol and Y. Fukushima for their contribution in analysing the ground deformations, F. Paganelli, J. Biggs and J. Keller for discussions and K. Feigl and F. Sigmundsson for their comments. We acknowledge support from the US National Science Foundation under grants EAR-0801801 and EAR-0538119, from the French INSU-CNRS DyETI programme and from the Belgian Science Policy under projects SAMAAV and Rukwa.

Author Contributions N.d'O., A.O. and F.K. planned the radar data acquisition and computed the interferograms; J.A., A.D., J.D., R.W.F. and J.P. designed the seismic experiment and collected and analysed the resulting data; D.D., A.S.M., B.S. and C.W. performed the field observations; E.S. and D.S.S. deployed the GPS equipment and processed the resulting data; C.E. provided tectonic context; E.C. performed the model calculations; E.C. and N.d'O. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.C. (ecalais@purdue.edu).

METHODS

Seismology. We determined the source mechanism of the 17 July $M_w = 5.9$ event using teleseismic body waves from 24 stations assuming mean crustal P- and S-wave velocities of 6.0 and $3.46\,\mathrm{km\,s^{-1}}$, respectively, and find a source of 3-s duration with a simple pulse, a seismic moment of 1.1×10^{18} N m and a purely normal faulting source mechanism with a strike of 241°, a dip of 55° and a rake of −85°. Earthquake locations were computed using the HYPOCENTER program[21] with a crustal velocity model derived from the 1991 seismic refraction experiment[22] of the Kenya Rift International Seismic Project, with P-wave velocities as given in Supplementary Table 1. Statistical uncertainties in hypocentre location are of 3–6 km in depth and north–south position and <3 km in east–west position.

Radar interferometry. We used radar data acquired by the ESA Envisat satellite[23]. Only images acquired on descending orbits are available for the area and time interval of interest, because of a failure in the Artemis satellite (ESA) relaying Envisat data to the ground. The main parameters from the two interferograms used here are given in Supplementary Table 2. Additional radar data from ALOS (a satellite of the Japanese Aerospace Exploration Agency) and Envisat are available; a complete catalogue can be found in ref. 4. Phase unwrapping was performed using the SNAPHU algorithm[24]. Satellite orbital parameters were fixed to the values provided by Delft University of Technology and ESA. We account for possible orbital biases in the slip inversion (see Methods) by solving for phase gradients in the north–south and east–west directions and a line-of-sight offset accounting for the unknown zero-phase level.

Modelling. We discretize the dyke and faults with rectangular patches of dimension ~1 km × 1 km and compute the impulse functions that relate slip on each patch[13] and volume change at each point source[14] to surface displacement at the observation points, while assuming a Poisson ratio of 0.25 and a shear modulus of 33 GPa. We use least squares to invert the model and estimate dyke-normal opening on the dyke patches, normal dip-slip displacement on the fault patches, and volume changes at the Mogi sources from InSAR line-of-sight range change and three-dimensional GPS displacement at site Engaresero.

We subsampled the interferograms to 302 data points for the interval between 3 April 2007 and 17 July 2007 and 1,254 data points for the interval between 17 July 2007 and 21 August 2007. The data were weighted by the inverse of their (diagonal) covariance matrix formed using a single-pixel standard deviation of 5.6 mm (20% of the half radar wavelength for Envisat), which was derived from the phase variations calculated from 20 interferograms spanning January 2006 to 2008 along two profiles, one across the southern flank of Gelai and the other in the Natron basin to the west of Gelai. To avoid implausible and overly rough slip/opening distributions, we applied smoothing by means of a finite-difference approximation of the Laplacian operator and applied positivity constraints.

Chi-squared statistics for the best-fit models are 98,832 and 148,455 and average residuals are 2.8 cm and 5.4 cm for the 3 April 2007–17 July 2007 and 17 July 2007–21 August 2007 intervals, respectively. We have verified that the simple elastic models used here provide results similar to forward models based on a three-dimensional mixed-boundary-element method that takes into account realistic topography and gravitational stresses[25]. Supplementary Fig. 1 compares, side-by-side, the observed unwrapped interferogram (left panel) with the simulated unwrapped interferogram derived from the estimated slip/opening model parameters (middle panel), and shows the residual signal (observed minus model).

Volume-change calculations depend strongly on Poisson's ratio (a constant value of 0.25 is used here), the shape of the magma chamber (here a point source) and magma degassing and expansion as pressure drops (not accounted for here)[26,27]. This later process may result in a volume increase by a factor of up to five[28].

21. Lienert, B. R. E. & Havskov, J. A computer program for locating earthquakes both locally and globally. *Seismol. Res. Lett.* **66,** 26–36 (1995).
22. KRISP Working Party. Large-scale variation in lithospheric structure along and across the Kenya rift. *Nature* **354,** 223–227 (1991).
23. Massonnet, D. & Feigl, K. L. Radar interferometry and its application to changes in the Earth's surface. *Rev. Geophys.* **36,** 441–500 (1998).
24. Chen, C. W. & Zebker, H. A. Two-dimensional phase unwrapping with use of statistical models for cost functions in nonlinear optimization. *J. Opt. Soc. Am. A* **18,** 338–351 (2001).
25. Cayol, V. & Cornet, F. H. Three-dimensional modeling of the 1983-1984 eruption at Piton de la Fournaise volcano, Réunion Island. *J. Geophys. Res.* **103,** 18025–18037 (1998).
26. Delaney, P. T. & McTigue, D. F. Volume of magma accumulation or withdrawal estimated from surface uplift or subsidence, with application to the 1960 collapse of Kilauea volcano. *Bull. Volcanol.* **56,** 417–424 (1994).
27. Johnson, D. J., Sigmundsson, F. & Delaney, P. T. Comment on ''Volume of magma accumulation or withdrawal estimated from surface uplift or subsidence, with application to the 1960 collapse of Kilauea volcano'' by P. T. Delaney and D. F. McTigue. *Bull. Volcanol.* **61,** 491–493 (2000).
28. Nishimura, T. Pressure recovery in magma due to bubble growth. *Geophys. Res. Lett.* **31,** doi:10.1029/2004GL019810 (2004).

LETTERS

Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic

Hélène Agogué[1], Maaike Brink[1], Julie Dinasquet[1] & Gerhard J. Herndl[1]†

Aerobic nitrification of ammonia to nitrite and nitrate is a key process in the oceanic nitrogen cycling mediated by prokaryotes[1]. Apart from Bacteria belonging to the β- and γ-Proteobacteria involved in the first nitrification step, Crenarchaeota have recently been recognized as main drivers of the oxidation of ammonia to nitrite in soil as well as in the ocean, as indicated by the dominance of archaeal ammonia monooxygenase (*amoA*) genes over bacterial *amoA*[2,3]. Evidence is accumulating that archaeal *amoA* genes are common in a wide range of marine systems[3–6]. Essentially, all these reports focused on surface and mesopelagic (200–1,000 m depth) waters, where ammonia concentrations are higher than in waters below 1,000 m depth. However, Crenarchaeota are also abundant in the water column below 1,000 m, where ammonia concentrations are extremely low. Here we show that, throughout the North Atlantic Ocean, the abundance of archaeal *amoA* genes decreases markedly from subsurface waters to 4,000 m depth, and from subpolar to equatorial deep waters, leading to pronounced vertical and latitudinal gradients in the ratio of archaeal *amoA* to crenarchaeal 16S ribosomal RNA (rRNA) genes. The lack of significant copy numbers of *amoA* genes and the very low fixation rates of dark carbon dioxide in the bathypelagic North Atlantic suggest that most bathypelagic Crenarchaeota are not autotrophic ammonia oxidizers: most likely, they utilize organic matter and hence live heterotrophically.

Planktonic Archaea consisting of the two major groups, Crenarchaeota and Euryarchaeota, might account for about one-third of all prokaryotic cells in the global ocean[7,8]. Crenarchaeota are generally more abundant than Euryarchaeota in marine waters[7–9] and mainly comprise the marine Crenarchaeota group I (MCGI) and the deep-branching pSL12-like clade[4]. Until recently, only members the bacterial domain, γ- and β-Proteobacteria were known to oxidize NH_4^+ to NO_2^-, and γ-, β-, δ-Proteobacteria and *Nitrospira* species to oxidize the intermediate product NO_2^- to NO_3^- (ref. 1). Recent genomic studies suggest that mesophilic Crenarchaeota might play a role in the ammonia oxidation, as the *amoA* gene encoding a subunit of one of the key enzymes in the ammonia oxidation, ammonia monooxygenase, has been detected[2,10,11]. The only two mesophilic marine crenarchaeal isolates (that is, *Cenarchaeum symbiosum* and *Nitrosopumilus maritimus*) also carry this *amoA* gene[12,13]. Recent quantitative studies showed that MCGI oxidize ammonia as their energy source and that archaeal *amoA* genes are at least one order of magnitude more abundant than bacterial *amoA* in various marine and soil environments[3–5,14,15]. Moreover, Mincer *et al.*[4] recently suggested that the pSL12-like clade may also contain *amoA* genes. Taken together, there is evidence that at least some of the MCGI are chemoautotrophic ammonia oxidizers although other studies indicate that Crenarcheaota might be heterotrophic as well[8,16–19].

All the studies quantifying archaeal *amoA* gene abundance in the pelagic realm, however, focused on the euphotic and mesopelagic zones, where ammonia is generated by remineralization processes such as in the oxygen minimum zone[3–5,11]. In addition, the only-free living isolate, *Nitrosopumilus maritimus*, has been obtained from an ammonia-rich tropical aquarium[12]. However, Crenarchaeota are abundant in bathypelagic (1,000–4,000 m depth) waters as well[20,21], where ammonia concentrations are below 5 nM (M. Woodward, personal communication) and only detectable with highly sensitive methods, thus making it unlikely that ammonia is readily available as an energy source in these deep waters.

The contributions of putatively ammonia-oxidizing Archaea (AOA) and Bacteria (AOB) to the total archaeal and bacterial communities were determined by quantitative PCR (qPCR) in the main water masses of the Atlantic from 65° N to 5° S covering a depth range from 100 m to 4,000 m. The abundance of specific groups of planktonic Crenarchaeota (MCGI and the pSL12-like clade) based on 16S rRNA gene fragments was determined and compared with the abundance of AOA and AOB using two specific primer sets targeting archaeal and bacterial *amoA* genes at 10 stations, comprising different water masses with specific physical and chemical characteristics (for details see Supplementary Fig. 1 and Supplementary Tables 1 and 2).

The highest MCGI abundance was found in the mesopelagic layer in almost all the stations, particularly in the oxygen minimum layer (that is, between 200 and 600 m depth; Fig. 1a and Supplementary Tables 1 and 3). The abundance of the pSL12-like clade was about 1.5 orders of magnitude lower than MCGI (Supplementary Table 3). In most of the samples (40 out of 55), the contribution of the pSL12-like clade to total crenarchaeal abundance was less than 5% (Supplementary Table 3) but followed a similar vertical distribution pattern as MCGI. The pSL12-like clade was more abundant just below the euphotic zone (that is, 250 m depth) than in deeper water masses (Supplementary Table 3). A similar trend was observed by Mincer *et al.*[4], who found more pSL12-related members below the euphotic zone of the North Pacific subtropical gyre. The distribution of the copy numbers of the 16S rRNA genes (MCGI and pSL12-clade) obtained by qPCR corresponds to the distribution pattern of Crenarchaeota with depths reported previously using fluorescent *in situ* hybridization[7,8,20,21]. Overall, marine Crenarchaeota seem to be particularly associated with the oxygen minimum zone[4,5,11,22].

Copy numbers of archaeal *amoA* genes were always highest at the base of the euphotic zone, ranging from 2.5×10^3 to 25×10^3 ml^{-1} and decreasing with depth (Fig. 1b and Supplementary Table 3). Although no latitudinal trend in the abundance of the archaeal *amoA* genes was detectable at the base of the euphotic zone, in deeper waters, archaeal *amoA* copy numbers decreased by two to three orders of magnitude from the subpolar to the subtropical and equatorial Atlantic (Fig. 1b and Supplementary Table 3). The higher abundance of archaeal *amoA* genes in the meso- and bathypelagic waters in the northern part of the Atlantic coincides with the higher

[1]Department of Biological Oceanography, Royal Netherlands Institute for Sea Research (Royal NIOZ), PO Box 59, 1790 AB Den Burg, Texel, The Netherlands. †Present address: Department of Marine Biology, Vienna Ecology Center, University of Vienna, 1090 Vienna, Austria.

ammonia concentrations there[21], as these deep waters have been recently formed and therefore are younger than in the southern part of the North Atlantic[23]. In contrast to the high number of archaeal *amoA* copy numbers, β-proteobacterial *amoA* copy numbers were at least one order of magnitude lower than archaeal *amoA* genes in the individual depth layers (Supplementary Table 3). γ-Proteobacterial *amoA* genes were not quantified in this study because ammonia-oxidizing γ-Proteobacteria were not detected in this region of the Atlantic by qPCR in a previous study[3] nor in Arctic or North Pacific samples[4,24]. Taken together, our data confirm previous findings that crenarchaeal *amoA* genes are far more abundant than bacterial *amoA* genes in marine systems[3,5,15].

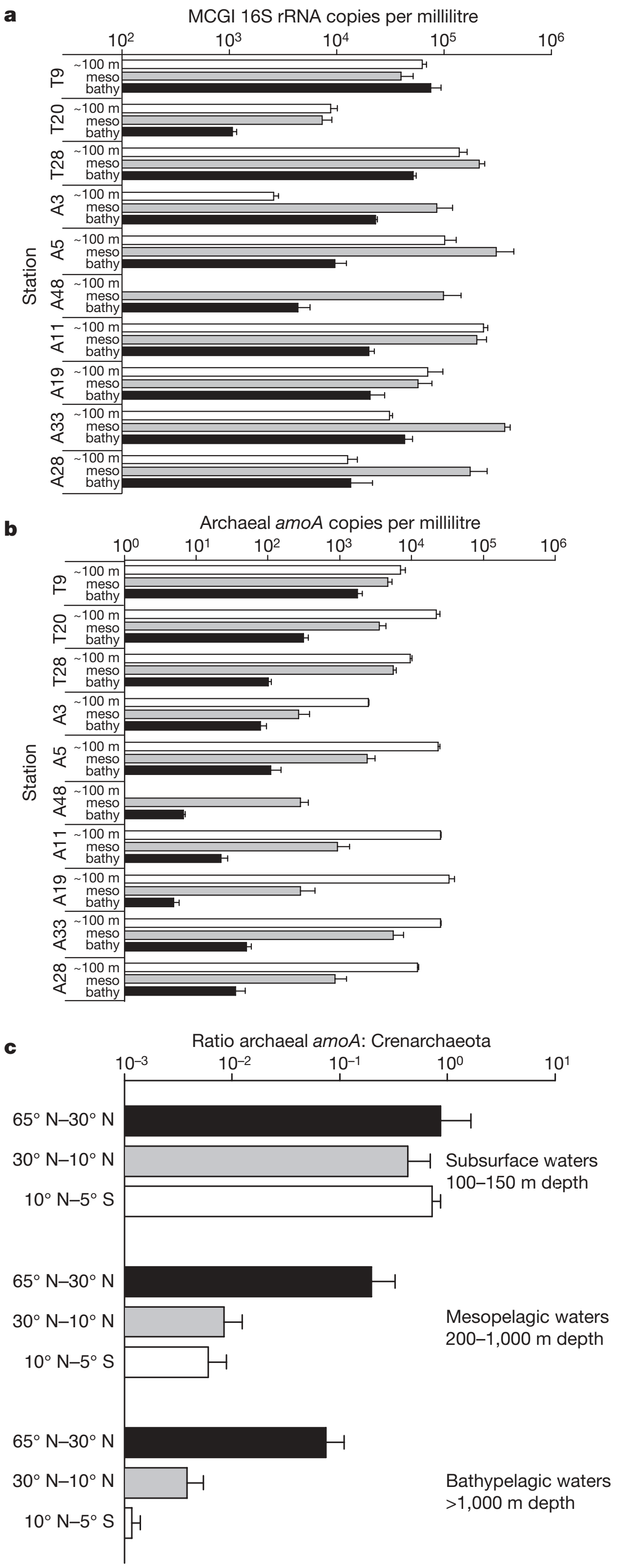


Figure 1 | Latitudinal and depth distribution of crenarchaeal 16S rRNA and archaeal *amoA* copy numbers in the eastern North Atlantic. **a**, **b**, Abundance of MCGI (**a**) and AOA (**b**) in the subsurface (~100 m) ($n = 9$), mesopelagic (meso) ($n = 24$) and bathypelagic (bathy) ($n = 22$) waters. **c**, Ratio of archaeal *amoA* copies to crenarchaeal (MCGI + pSL12-like clade) 16S rRNA copies per millilitre of marine water ($n = 55$). Bars indicate standard errors.

The average ratio of archaeal *amoA* copy numbers compared with crenarchaeal 16S rRNA genes in the subsurface waters (100–150 m depth) was close to 1, particularly in the northern and equatorial North Atlantic, whereas it was less than 0.01 in the bathypelagic waters of the subtropical and equatorial region of the North Atlantic (Fig. 1c and Supplementary Fig. 2). Our ratios for the surface waters are similar to those of Mincer *et al.*[4], who reported a ratio between archaeal *amoA* and archaeal 16S rRNA genes of about 1, whereas Wuchter *et al.*[3] and Beman *et al.*[5] reported a ratio of 2.8 and 2.5, respectively. Genomic studies on the two mesophilic crenarchaeal isolates obtained thus far revealed that *Cenarchaeum symbiosum* and *Nitrosopumilus maritimus* contain only one *amoA* gene copy per cell[13]. In the 100–150 m depth range, the average ratios between archaeal *amoA* and crenarchaeal 16S rRNA genes were 0.9, 0.4 and 0.7 for the northern, subtropical and equatorial regions, respectively (Fig. 1c). We conclude, therefore, that in subsurface waters most of the Crenarchaeota are putatively capable of ammonia oxidation throughout the North Atlantic. In the northern part of the Atlantic (65–30° N), the ratio of archaeal *amoA* to crenarchaeal 16S rRNA genes decreased only by about one order of magnitude from the surface waters to the bathypelagic realm (Fig. 1c). In the subtropical and equatorial regions of the Atlantic (30° N – 5° S), however, this ratio decreased by at least two orders of magnitude from the subsurface to the bathypelagic waters (Fig. 1c). Hence, AOA are apparently only abundant throughout the water column in the northern North Atlantic, with its newly formed and therefore young deep waters[23,25,26]. The major fraction of the abundant Crenarchaeota in the subtropical and tropical North Atlantic deep waters most likely use energy sources other than ammonia.

To assess the diversity of AOA in the meso- and bathypelagic waters of the North Atlantic, archaeal *amoA* clone libraries were established for two mesopelagic samples where qPCR analyses revealed relatively high numbers of AOA, and for one bathypelagic sample where the AOA abundance was very low (Supplementary Table 3). The bathypelagic clone library exhibited the lowest observed and estimated operational taxonomic unit richness (Supplementary Fig. 3 and Supplementary Table 4). The richness levels of the two mesopelagic clone libraries were comparable or higher than those observed for archaeal *amoA* clone libraries from other marine pelagic environments[11] (Supplementary Table 4). In a DNA-based phylogenetic tree, the obtained sequences fell into two distinct groups: cluster A ('shallow' cluster) and cluster B ('deep' cluster)[5,11,27] (Fig. 2). Based on phylogenetic analyses of PCR clone libraries, individual regions of the mesopelagic North Atlantic harbour specific archaeal *amoA* sequences (that is, North Atlantic and equatorial subclusters (Fig. 2)). More than half of the sequences of the clone library from the mesopelagic North Atlantic formed a coherent cluster closely related to sequences obtained from the same region[3]. Similarly, specific sequences recovered from the bathypelagic clone library fell into bathypelagic subcluster 1 within cluster A and into bathypelagic subcluster 2 of the phylogenetically distinct cluster B (Fig. 2). About 60% of the obtained archaeal *amoA* sequences were affiliated to the North Atlantic subcluster harbouring *amoA* sequences from a diverse range of marine habitats[3–5,11,22] (Fig. 2). Taken together, our phylogenetic analyses indicate latitudinal and vertical differences among archaeal *amoA* sequences with generally rather low diversity in equatorial bathypelagic waters (Supplementary Fig. 3) coinciding with the rather low abundance of AOA.

There is, however, still the possibility that deep ocean Crenarchaeota harbour specific versions of *amoA* genes not targeted by the available primer sets.

An archaeal 16S rRNA clone library was also established for one mesopelagic sample and three bathypelagic samples (for details see Supplementary Information). In a DNA-based phylogenetic tree, the obtained crenarchaeal sequences fell into the MCGI and pSL12-like groups (Supplementary Fig. 4). The phylogenetic analyses of the archaeal 16S rRNA genes also strengthen the idea that specific water masses harbour specific ecotypes and that the presence of operational taxonomic unit 2, closely related to *Nitrosopumilus maritimus*, especially in the TRANSAT sample, might explain the high abundance of *amoA* genes in this specific water mass (for details see Supplementary Information).

Ammonia-oxidizing Archaea are generally chemoautotrophic, using acetyl-coenzyme A carboxylase to fix inorganic carbon[28].

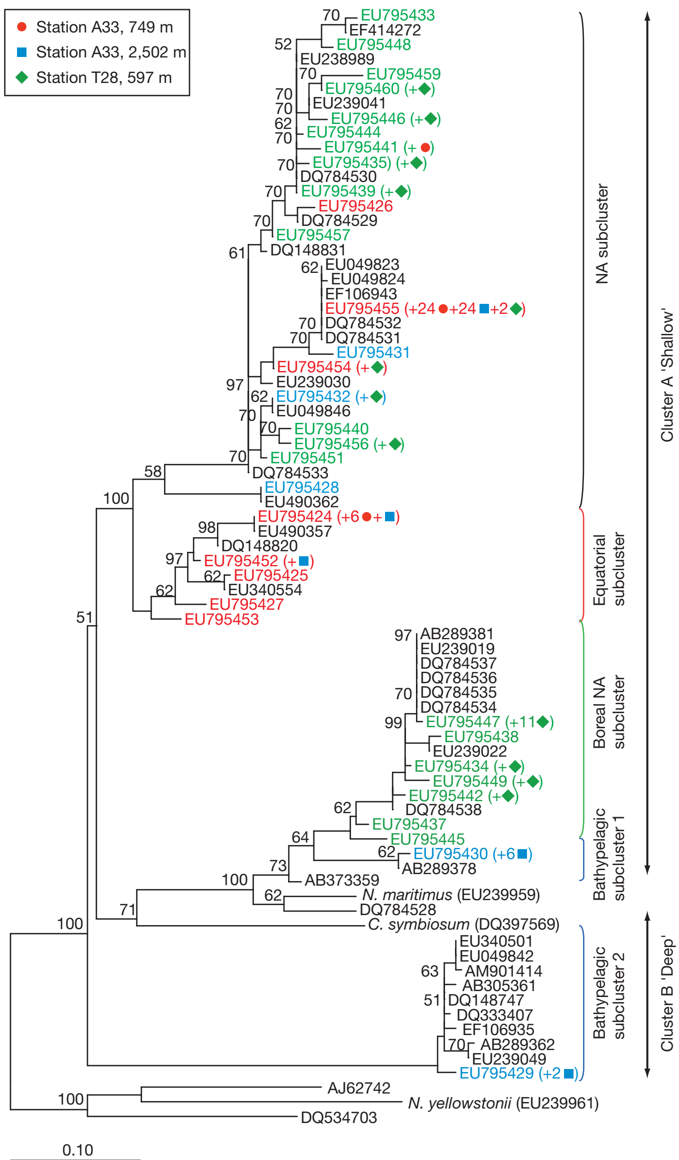


Figure 2 | Phylogenetic tree of archaeal *amoA* sequences recovered from the North Atlantic. Neighbour-joining bootstrap tree of long *amoA*-like nucleotide sequences to which we added our partial sequences (217 base pairs (bp), EU795424–EU795460) and the sequences of Wuchter *et al.*[3] (217 bp, DQ784528–DQ784537) using the ARB parsimony tool (NA, North Atlantic). One representative of sequence group ≥99% identical is shown (in colour); the number in front of the respective symbol indicates the number of clones. Database sequences are shown in black. Bootstrap values (>50%) are indicated at each branch point.

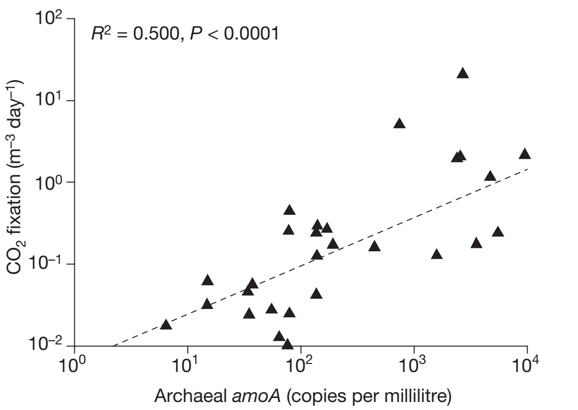


Figure 3 | Archaeal *amoA* copy numbers versus CO_2 fixation. Relation between archaeal *amoA* gene copy numbers and CO_2 fixation across all the stations and depths (except station T9; $n = 31$). Regression equation: $y = 0.0063x^{0.5905}$.

Hence, if AOA are the dominating nitrifiers, as suggested by the much higher archaeal than bacterial *amoA* copy numbers, one would expect a tight relation between archaeal *amoA* copy numbers and bulk CO_2 fixation rates. Relating the dark CO_2 fixation rates obtained at these specific stations and depths by Herndl *et al.*[8] and T. Reinthaler and G.J.H. (unpublished observations) to the archaeal *amoA* copy numbers determined in this study, it appears that about 50% of the variation in CO_2 fixation rates can be explained by archaeal *amoA* abundance (Fig. 3). This provides further independent evidence that most bathypelagic Crenarchaeota are not chemoautotrophic nitrifiers, but most likely use organic substrates as carbon and energy sources, as also indicated by microautoradiography combined with fluorescence *in situ* hybridization and genomic analyses[8,16–18,21,27]. This conclusion is further supported by recent comparative metagenomic analyses of uncultivated Archaea of various meso- and bathypelagic oceanic regions by Martin-Cuadrado *et al.*[29]. In none of the nine fosmid libraries of MCGI from 500 to 3,000 m depth were *amoA* genes detected[29].

In summary, it appears that only in the northern part of the North Atlantic mesophilic Archaea predominantly oxidize ammonia as an energy source throughout the water column including the bathypelagic realm. The abundance of putatively ammonia-oxidizing Crenarchaeota gradually decreases in the meso- and bathypelagic waters from the north to the equator in the Atlantic Ocean, coinciding with an increasing age of these water masses and, concomitantly, a decrease in the diversity of archaeal *amoA* sequences. In the subtropical and equatorial bathypelagic waters, AOA contribute less than 1% to the crenarchaeal community. Hence, only a minor fraction of the bathypelagic Crenarchaeota putatively oxidize ammonia in the temperate and subtropical North Atlantic. Consequently, a heterotrophic life mode for bathypelagic Crenarchaeota seems to become increasingly important from north to south in the Atlantic's interior.

METHODS SUMMARY

Sampling was conducted aboard the RV *Pelagia* during two cruises (TRANSAT-1 and ARCHIMEDES-2) following the North Atlantic Deep Water from 65° N to 5° S in the eastern basin of the Atlantic Ocean. Specific water masses were identified and sampled based on their physical and chemical characteristics.

Using qPCR, the copy numbers of MCGI 16S rRNA, pSL12-like 16S rRNA, archaeal *amoA* and β-proteobacterial *amoA* genes were determined using specific primers and SYBR Green.

Archaeal *amoA* clone libraries were generated from two stations: station A33 at two distinct depths (749 and 2,502 m) and station T28 at 597 m depth by PCR amplification using specific archaeal *amoA* primers. Archaeal 16S rRNA gene clone libraries were generated from three stations: station T9 at 2,071 m, station A3 at two distinct depths (249 and 2,750 m) and one additional station in the Romanche

Fracture Zone (-0.21° N, -18.43° E) near station A33 (1° N, -20.60° E) at 7,155 m depth by PCR amplification using specific archaeal primers.

Dark archaeal CO_2 fixation was measured by the incorporation of [^{14}C]bicarbonate. The resulting mean disintegrations per minute of the samples were converted into organic carbon produced over time and corrected for the natural dissolved inorganic carbon concentration.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 September; accepted 13 October 2008.
Published online 26 November 2008.

1. Ward, B. B., Capone, D. G. & Zehr, J. P. What's new in the nitrogen cycle? *Oceanography* **20,** 101–109 (2007).
2. Schleper, C., Jurgens, G. & Jonuscheit, M. Genomic studies of uncultivated Archaea. *Nature Rev. Microbiol.* **3,** 479–488 (2005).
3. Wuchter, C. *et al.* Archaeal nitrification in the ocean. *Proc. Natl Acad. Sci. USA* **103,** 12317–12322 (2006).
4. Mincer, T. J. *et al.* Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environ. Microbiol.* **9,** 1162–1175 (2007).
5. Beman, J. M., Popp, B. N. & Francis, C. A. Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. *ISME J.* **2,** 429–441 (2008).
6. Caffrey, J. M., Bano, N., Kalanetra, K. & Hollibaugh, J. T. Ammonia oxidation and ammonia-oxidizing bacteria and archaea from estuaries with differing histories of hypoxia. *ISME J.* **1,** 660–662 (2007).
7. Karner, M. B., DeLong, E. F. & Karl, D. M. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409,** 507–510 (2001).
8. Herndl, G. J. *et al.* Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. *Appl. Environ. Microbiol.* **71,** 2303–2309 (2005).
9. Church, M. J. *et al.* Abundance and distribution of planktonic Archaea and Bacteria in the waters west of the Antarctic Peninsula. *Limnol. Oceanogr.* **48,** 1893–1902 (2003).
10. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
11. Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E. & Oakley, B. B. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc. Natl Acad. Sci. USA* **102,** 14683–14688 (2005).
12. Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437,** 543–546 (2005).
13. Hallam, S. J. *et al.* Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum. Proc. Natl Acad. Sci. USA* **103,** 18296–18301 (2006).
14. Leininger, S. *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442,** 806–809 (2006).
15. Nakagawa, T., Mori, K., Kato, C., Takahashi, R. & Tokuyama, T. Distribution of the cold-adapted ammonia-oxidizing microorganisms in the deep-ocean of the northeastern Japan Sea. *Microbes Environ.* **22,** 365–372 (2007).
16. Ouverney, C. C. & Fuhrman, J. A. Marine planktonic Archaea take up amino acids. *Appl. Environ. Microbiol.* **66,** 4829–4833 (2000).
17. Teira, E., van Aken, H., Veth, C. & Herndl, G. J. Archaeal uptake of enantiomeric amino acids in the meso- and bathypelagic waters of the North Atlantic. *Limnol. Oceanogr.* **51,** 60–69 (2006).
18. Kirchman, D. L., Elifantz, H., Dittel, A. I., Malmstrom, R. R. & Cottrell, M. T. Standing stocks and activity of Archaea and Bacteria in the western Arctic Ocean. *Limnol. Oceanogr.* **52,** 495–507 (2007).
19. Ingalls, A. E. *et al.* Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc. Natl Acad. Sci. USA* **103,** 6442–6447 (2006).
20. Teira, E., Lebaron, P., van Aken, H. & Herndl, G. J. Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic. *Limnol. Oceanogr.* **51,** 2131–2144 (2006).
21. Varela, M. M., van Aken, H. M., Sintes, E. & Herndl, G. J. Latitudinal trends of *Crenarchaeota* and *Bacteria* in the meso- and bathypelagic water masses of the Eastern North Atlantic. *Environ. Microbiol.* **10,** 110–124 (2008).
22. Lam, P. *et al.* Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea. *Proc. Natl Acad. Sci. USA* **104,** 7104–7109 (2007).
23. Rhein, M. *et al.* Labrador Sea Water: pathways, CFC inventory, and formation rates. *J. Phys. Oceanogr.* **32,** 648–665 (2002).
24. Hollibaugh, J. T., Bano, N. & Ducklow, H. W. Widespread distribution in polar oceans of a 16S rRNA gene sequence with affinity to *Nitrosospira*-like ammonia-oxidizing bacteria. *Appl. Environ. Microbiol.* **68,** 1478–1484 (2002).
25. Smethie, W. M. & Fine, R. A. Rates of North Atlantic Deep Water formation calculated from chlorofluorocarbon inventories. *Deep-Sea Res. I* **48,** 189–215 (2001).
26. van Aken, H. M. The hydrography of the mid-latitude northeast Atlantic Ocean: I: the deep water masses. *Deep-Sea Res. I* **47,** 757–788 (2000).
27. Hallam, S. J. *et al.* Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota. PLoS Biol.* **4,** 520–536 (2006).
28. Berg, I. A., Kockelkorn, D., Buckel, W. & Fuchs, G. A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in archaea. *Science* **318,** 1782–1786 (2007).
29. Martin-Cuadrado, A.-B. *et al.* Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.* **2,** 865–886 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the captain and crew of the RV *Pelagia* for their help during work at sea. We thank H. M. van Aken for the characterization of the water masses, J. M. Arrieta for collecting the samples during the TRANSAT-1 cruise, B. Abbas, J. van Bleijswijk and H. Witte for technical discussions about the qPCR and phylogenetic analyses, and T. Reinthaler for discussions. We also thank A. Hunting, A. M. Schmitz and W. D. Lienhart for help with data processing. This research was supported by a Marie Curie Fellowship of the European Community to H.A. Shiptime was provided through grants of the Earth and Life Science Division of the Dutch Science Foundation (ALW-NWO) (TRANSAT and ARCHIMEDES projects) to G.J.H. The work was performed within the frame of the 'Networks of Excellence' MarBef and EurOceans supported by the 6th Framework Program of the European Union.

Author Contributions The manuscript was written by H.A. and G.J.H. DNA extractions were performed by H.A. and M.B., qPCR by H.A., M.B. and J.D. and phylogenetic analyses by M.B. Measurements of dissolved inorganic carbon incorporation were done by G.J.H.

Author Information The archaeal sequences are deposited in GenBank under accession numbers EU650236–EU650270 (station 3, ARCHIMEDES-2), FJ002858–FJ002876 (station 23, ARCHIMEDES-3), FJ150794–FJ150834 (Station 9, TRANSAT-1) for 16S rRNA genes and EU795424–EU795460 and EU810209–EU810235 for *amoA* genes. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.J.H. (herndl@nioz.nl).

METHODS

Study sites and sampling. Sampling was conducted aboard the RV *Pelagia* during two cruises (TRANSAT-1 and ARCHIMEDES-2) following the North Atlantic Deep Water from 65° N to 5° S in the eastern basin of the Atlantic Ocean. Specific water masses were identified and sampled based on their physical and chemical characteristics (Supplementary Table 1). At each station, six depths were sampled with 10-l NOEX samplers mounted on a conductivity–temperature–depth (CTD) frame. Ten stations were used for qPCR analysis (Supplementary Fig. 1).

DNA extraction of marine samples. Extraction of total DNA was performed using an UltraClean soil DNA and Mega soil DNA isolation kits (Mobio) for TRANSAT-1 and ARCHIMEDES-2 samples, respectively.

qPCR. The copy numbers of MCGI 16S rRNA, pSL12-like 16S rRNA, archaeal *amoA* and β-proteobacterial *amoA* genes were determined in triplicate on the non-diluted samples and in two different dilutions per sample using specific primers (Supplementary Table 2) and SYBR Green. The four real-time approaches yielded highly reproducible standard curves with DNA of the plasmid 088exp4, the fosmid HF770_041/11, *Nitrosopumilus maritimus* and a mix of four β-Proteobacteria, for MCGI, pSL12-like, archaeal *amoA* and β-proteobacterial *amoA* standards, respectively. All reactions were performed in duplicate for the standards and in triplicate for the samples; an average value was calculated for each sample. Two different methods were used to determine the presence of non-specific amplified products. No unspecific products were found in any of the analysis (see Supplementary Information for details).

Archaeal *amoA* clone libraries. Archaeal *amoA* clone libraries were generated from two stations: station A33 at two distinct depths (749 and 2,502 m) and station T28 at 597 m depth by PCR amplification using specific archaeal *amoA* primers (Supplementary Fig. 1 and Supplementary Table 2). The obtained sequences were compared with sequences from public databases and analysed using the ARB phylogenetic package (see Supplementary Information for details).

Archaeal 16S rRNA gene clone libraries. Archaeal 16S rRNA gene clone libraries were generated from three stations: station T9 at 2,071 m, station A3 at two distinct depths (249 and 2,750 m) and one additional station in the Romanche Fracture Zone (−0.21° N, −18.43° E) near station A33 (1° N, −20.60° E) at a depth of 7,155 m by PCR amplification using specific archaeal primers (Supplementary Fig. 1 and Supplementary Table 2). The obtained sequences were compared with sequences from public databases and analysed using the ARB phylogenetic package (see Supplementary Information for details).

Archaeal CO_2 fixation. Dark CO_2 fixation was measured by the incorporation of [^{14}C]bicarbonate (100 μCi, Amersham) in 40-ml samples in triplicate and triplicate formaldehyde-fixed blanks incubated in the dark at *in situ* temperature for 60–72 h. Subsequently, the incubations were terminated by adding formaldehyde (2% final concentration) to the samples, filtering onto 0.2-μm filters (Millipore, polycarbonate), rinsing three times with 10 ml of ultrafiltered water (30 kDa molecular mass cut off) and then exposing the filters to a fume of concentrated HCl for 12 h. Thereafter, the filters were placed in scintillation vials, 8 ml of scintillation cocktail (FilterCount, Canberra-Packard) was added and, after about 18 h, counted on board in a liquid scintillation counter (LKB Wallac). The instrument was calibrated with internal and external standards. The resulting mean numbers of disintegrations per minute of the samples were corrected for the mean disintegrations per minute of the blanks, converted into organic carbon produced over time and corrected for the concentration of natural dissolved inorganic carbon.

CORRIGENDUM

doi:10.1038/nature07735

Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic

Hélène Agogué, Maaike Brink, Julie Dinasquet & Gerhard J. Herndl

Nature 456, 788–791 (2008)

The *y* axis in Fig. 3 of this Letter was incorrectly labelled 'CO_2 fixation (m^{-3} day^{-1})'. It should read 'CO_2 fixation (μmol C m^{-3} day^{-1})'.

Vol 456 | 11 December 2008 | doi:10.1038/nature07395

LETTERS

Strong effect of dispersal network structure on ecological dynamics

Matthew D. Holland[1] & Alan Hastings[1]

A central question in ecology with great importance for management, conservation and biological control is how changing connectivity affects the persistence and dynamics of interacting species. Researchers in many disciplines have used large systems of coupled oscillators to model the behaviour of a diverse array of fluctuating systems in nature[1–4]. In the well-studied regime of weak coupling, synchronization is favoured by increases in coupling strength and large-scale network structures (for example 'small worlds') that produce short cuts and clustering[5–9]. Here we show that, by contrast, randomizing the structure of dispersal networks in a model of predators and prey tends to favour asynchrony and prolonged transient dynamics, with resulting effects on the amplitudes of population fluctuations. Our results focus on synchronization and dynamics of clusters in models, and on timescales, more appropriate for ecology, namely smaller systems with strong interactions outside the weak-coupling regime, rather than the better-studied cases of large, weakly coupled systems. In these smaller systems, the dynamics of transients and the effects of changes in connectivity can be well understood using a set of methods including numerical reconstructions of phase dynamics, examinations of cluster formation and the consideration of important aspects of cyclic dynamics, such as amplitude.

Our study of the role of network structure on ecological dynamics considers exploiter–victim interactions[10–14] in networks that are relatively small and heterogeneous, and in which the importance of the amplitude of cycles means that results from weak coupling do not apply. We use explicit models of ecological dynamics, in which issues concerning both the phases and the amplitudes of oscillators are important, as opposed to general models based on phase oscillators[5,15]. The systems of conservation or ecological interest are typically those with strong connections[16], where boom-and-bust cycles can bring interacting populations close to extinction. Previous work suggests the importance of spatial structure and connectivity for reduction of cycle amplitude and increased persistence in both models[10–12,17,18] and laboratory systems[19]. More recently, great interest has arisen in the practical application of connectivity in a conservation context and in biological control[14], with all the heterogeneities present in natural and managed systems having a vital role[16,20–22]. However, most current theoretical understanding comes from models with very regular connections among patches, by contrast with the heterogeneity in natural systems that are far from regular[22–24]. We focus on heterogeneities, specifically in network structure, and on the formation of clusters, that is, groups of spatial locations with synchronous dynamics.

We present results for ten-patch systems, which are intermediate in size between typical studies of small systems (two to four patches[17,25]) and large lattice systems (hundreds to thousands of patches[4,12]). We choose this size mainly for tractability, and also because many managed systems have a relatively small number of connected sites, and larger systems behave similarly. We use regular lattices of degree four (each node, which represents a habitat patch, is connected to four nearest neighbours) to begin modelling dispersal networks. To investigate the effect of irregular patterns of dispersal among patches, we also consider networks that have been randomized by 'rewiring' one or more edges[26] (removing a connection between one pair of patches and replacing it with a connection between a different pair), as well as networks in which all edges (connections) have been assigned at random. Sample networks are shown in Fig. 1. The local and dispersal dynamics of the system are modelled as follows:

$$\frac{\mathrm{d}h_i}{\mathrm{d}t} = h_i(1-\theta h_i) - \frac{p_i h_i}{1+h_i} + d_h \sum_{j=1}^{n} A_{ij} h_j \qquad (1)$$

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = \frac{\phi p_i h_i}{1+h_i} - \eta p_i + d_p \sum_{j=1}^{n} A_{ij} p_j \qquad (2)$$

This is a non-dimensional and spatially structured form of the well-known Rosenzweig–MacArthur predator–prey model[27], with h representing prey density and p representing predator density, interacting in n discrete habitat patches[25]. The local dynamics are determined by three parameters: ϕ (prey-to-predator conversion rate), θ (strength of prey self-regulation) and η (predator mortality rate). Dispersal among patches is governed by the dispersal rates d_h (prey) and d_p (predators) and the matrix A.

The system of equations (1) and (2) can produce a variety of spatial dynamics. Solutions range from global synchrony (all patches display identical fluctuations through time; we will refer to this as the one-cluster solution) to global asynchrony (all patches have unique trajectories at any given time; n-cluster solutions), as well as solutions with between two and $n-1$ clusters of synchronous patches (Fig. 2 and Supplementary Figs 1–6). It is important to note that (even for the same parameter values) there may be several k-cluster solutions with different characteristics for $2 \leq k \leq n$. In particular, continuous measures of asynchrony such as correlation of densities among

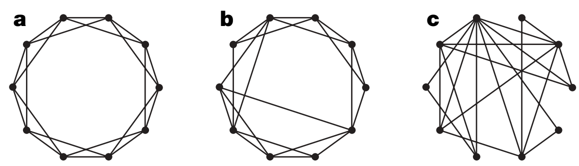


Figure 1 | Sample dispersal networks for systems with ten patches. All networks have an average degree of four, and dispersal is bidirectional. **a**, A ring lattice, referred to as a 'regular' network. **b**, A rewired ring lattice starts as a ring lattice and is rewired by replacing m edges at random. Here $m = 2$. **c**, A random network is produced by choosing $nk/2$ edges at random from all combinations of vertices (i, j) such that $i < j$. All dispersal networks used here were connected (every node can be reached by traversing one or more edges from every other node).

[1]Department of Environmental Science and Policy, University of California, Davis, One Shields Avenue, Davis, California 95616, USA.

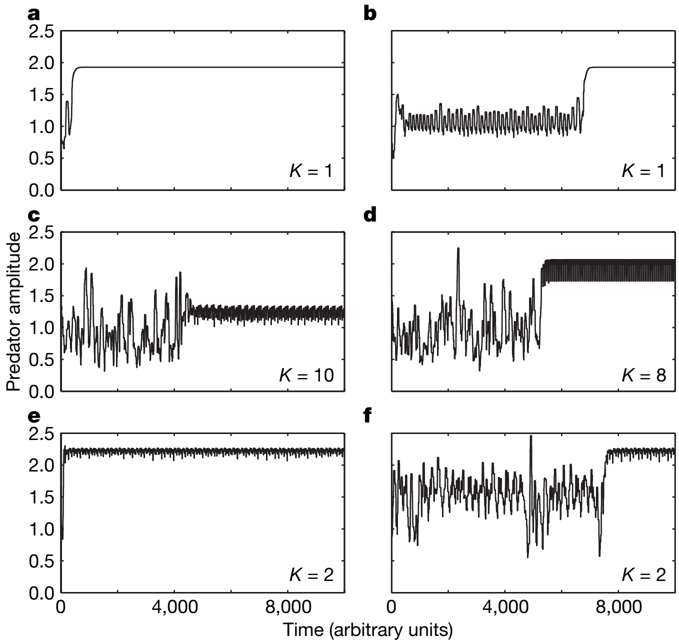


Figure 2 | Total predator amplitude (summed over all patches) as a function of time for cluster solutions. Local dynamics are characterized by strong prey density dependence ($\theta = 0.3$), predator mortality rates ($\eta = 1$) comparable to prey birth rates, moderate predator dispersal ($d_p = 2^{-7}$) and slower prey dispersal ($d_h = 2^{-9}$). **a**, **b**, Weak predation ($\phi = 2.75$) and regular networks. **c**–**f**, Strong predation ($\phi = 6$) and rewired networks ($m = 2$). All initial conditions and rewired networks were independently generated (see Methods). Each panel is labelled with the asymptotic number of clusters (K) observed in the simulation.

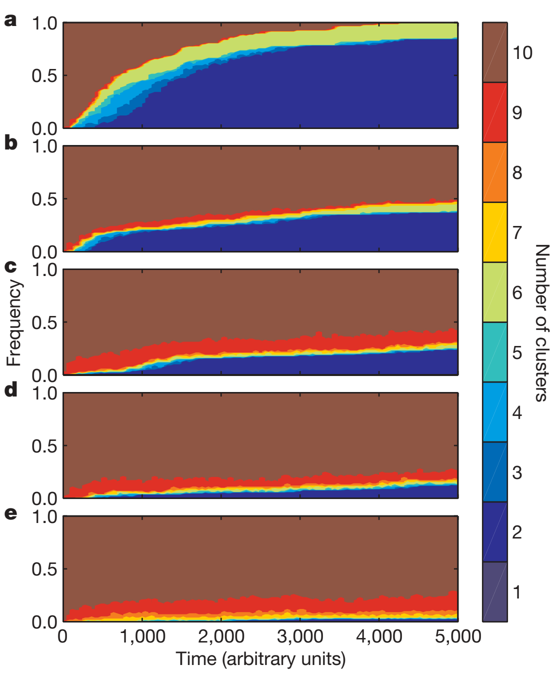


Figure 3 | Distribution of cluster states as a function of time. We consider systems with moderate dispersal rates ($d_h = 2^{-9}$, $d_p = 2^{-7}$) and ordered to random dispersal networks. **a**, Regular lattice; **b**, $m = 1$; **c**, $m = 2$; **d**, $m = 3$; **e**, random network. Regular networks converge to two- and six-cluster solutions, whereas irregular networks produce more eight-, nine- and ten-cluster solutions. Most of these large-k-cluster states are transient solutions. Local dynamics correspond to strong predation ($\phi = 5$), strongly self-regulating prey ($\theta = 0.3$) and predator mortality rates ($\eta = 1$) comparable to prey birth rates. Each panel summarizes 100 independent simulations.

patches in different clusters may vary among k-cluster solutions and over time for a single k-cluster solution. Thus, even 'globally asynchronous' n-cluster solutions may display considerable synchrony and, hence, larger amplitudes at various times during the solution. Furthermore, large-k-cluster classes can contain complex transient solutions resembling chaotic saddles, as well as asymptotic solutions displaying chaotic, quasi-periodic or periodic behaviour.

We find consistent and striking differences between dynamics with regular and irregular topologies for a broad range of underlying local dynamics. Looking at the distribution of cluster number through time in an ensemble of simulations reveals a substantial amount of information about the transient and asymptotic dynamics of these systems (Fig. 3 and Supplementary Figs 7–11). Increasing randomization of the network dramatically reduces the proportion of solutions with small to moderate numbers of clusters over ecologically relevant timescales.

For systems with heterogeneous dispersal networks, the temporal dynamics of k-cluster solutions suggest that transient dynamics are probably far more important than asymptotic dynamics on ecological timescales (Fig. 3). There are consequences for fluctuation amplitude and extinction risk of predators and prey, as is further revealed by manipulating both the local and the dispersal dynamics of the system. Varying predator efficiency, ϕ, is one way to move the system from low- to high-amplitude fluctuations. Because we are mainly interested in spatial persistence mechanisms, regional amplitude is a reasonable proxy for extinction risk in the context of a deterministic model (see Methods). For relatively low predator efficiencies, the asymptotic dynamics of all patch configurations are globally synchronous (Fig. 4a) and transients are short (Fig. 4b). At higher predator efficiencies, intermediate k-cluster solutions become stable, and asymptotic amplitudes level off near two orders of magnitude. This effect holds even at very high predator efficiencies, where single-patch systems fluctuate over five or more orders of magnitude and local extinction is extremely likely. The transient dynamics of these solutions have even lower median amplitudes, closer to one order of magnitude. Systems with irregular network structures spend much more time on these lower-amplitude transient solutions (Fig. 4b).

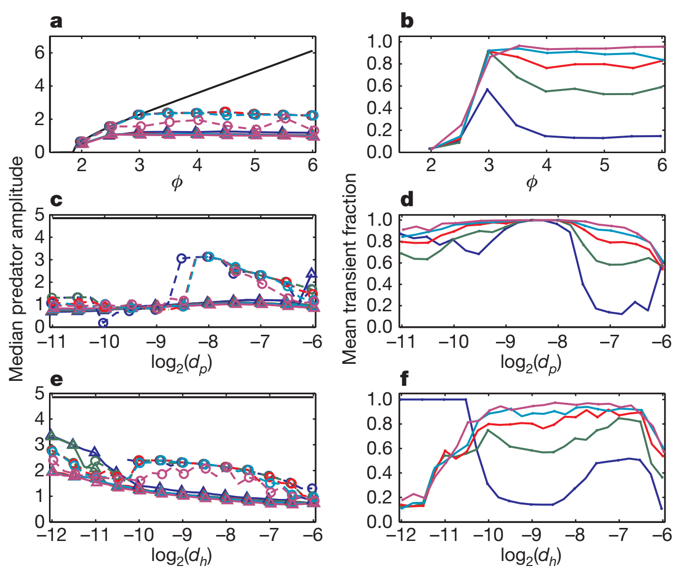


Figure 4 | Predator amplitude and transient duration. **a**, **c**, **e**, Median total predator amplitude during transient (triangles, solid lines) and asymptotic (open circles, dashed lines) solution phases; **b**, **d**, **f**, mean fraction of time spent in transient solutions. The black lines in **a**, **c**, and **e** are amplitudes of globally synchronous solutions for the chosen parameter values: $\theta = 0.3$, $\eta = 1$, $d_h = 2^{-9}$, $d_p = 2^{-7}$ (**a**, **b**); $\phi = 5$, $\theta = 0.3$, $\eta = 1$, $d_h = 2^{-9}$ (**c**, **d**), $\phi = 5$, $\theta = 0.3$, $\eta = 1$, $d_p = 2^{-7}$ (**e**, **f**). Data points are jittered horizontally and alternate markers omitted in **c**–**e** to improve readability. Each case uses at least 150 independent simulations. Dark blue, regular lattice; green, $m = 1$; red, $m = 2$; light blue, $m = 3$; pink, random network.

Manipulating dispersal rates leads to more complex spatio-temporal patterns. As predator dispersal is increased from a low value, we go from long transients and low-amplitude asymptotic solutions to a regime of higher amplitude asymptotic solutions preceded by lower-amplitude transients, with transient durations decreasing at higher movement rates (Fig. 4c, d). Very low values of prey dispersal produce shorter transients and low amplitudes in randomized networks and long transients of high amplitude in highly ordered lattices ($\log_2(d_h) < -10$; Fig. 4e, f). Larger rates of prey movement yield the now-familiar pattern of short transients leading to higher asymptotic amplitudes in ordered lattices, and longer transients of low amplitude in more disordered networks ($\log_2(d_h) > -10$, Fig. 4e, f).

Heterogeneous dispersal networks lead to dynamics with larger numbers of clusters (Fig. 3), which are generally transient solutions with considerable asynchrony and lower median amplitudes (Fig. 4) than the asymptotic solutions observed for lattice dispersal networks. There are some important caveats, however. Some transient solutions appear to be chaotic saddles or quasi-periodic solutions in which the median amplitude may not represent the dynamical regime that actually determines persistence. These solutions are relatively rare in our simulations, and occur most often for cases of strong resonance (Fig. 4c) or for very high values of predator dispersal (Supplementary Figs 14, 15). The relevance of such cases will of course depend on the details of the biological system being studied.

We demonstrate here that irregularities in connections among different sites in ecological networks can have large effects on the resulting dynamics. In particular, heterogeneous networks typically have longer periods of asynchronous dynamics, leading typically to lower amplitude fluctuations in population abundances; consequently, these are irregularities that cannot be ignored. However, decisions about the use of connectivity and corridors are being made in ecological systems based for the most part on theory that does not explicitly included heterogeneities[16,28]. Simultaneously, there is the need to characterize complex multidimensional systems in ecology, especially in the face of limited data. We show that for the kinds of systems we consider, cluster size is a simple, yet powerful and informative, way of characterizing dynamics, which both has a relationship to underlying network regularity and can help in understanding the dynamics of transient solutions on ecological timescales.

METHODS SUMMARY

The system of equations (1) and (2) was integrated numerically using a method appropriate for stiff equations. All simulations were initiated with pseudorandomly generated initial conditions, with all prey densities independently and identically distributed (i.i.d.) and all predator densities i.i.d. among all patches. For simulations using rewired or random networks, each simulation used an independently generated network.

Windowed analyses of cluster solutions (Fig. 3, Supplementary Figs 7–11) and amplitude (Figs 2, 4, Supplementary Figs 14, 15) were carried out on overlapping windows of length $4\bar{T}$ beginning at unit time intervals, where $\bar{T}$ is the mean period of a predator–prey cycle in a given simulation, averaged over all patches.

We defined a k-cluster solution as a solution in which the n patches in the system could be assigned to $k \leq n$ clusters of patches with identical dynamics. Clusters were taken to be the unions of all intersecting pairs of synchronous patches.

Time series were partitioned into transient and asymptotic regimes by treating predator–prey trajectories in individual patches as phase oscillators. We assumed that the rate of phase evolution within each patch would be constant when asymptotic dynamics had been reached. In fact, many of the asymptotic solutions contain periodic variation in the oscillator periods themselves. Thus, the transient phase of dynamics was defined as the interval of time before all patches displayed constant or periodic phase evolution. An automated algorithm (described in Methods) was used to estimate this value from numerical solutions to equations (1) and (2).

Total predator amplitude (Figs 2, 4) was defined simply as

$$\log_{10} \frac{\max\left(\sum_{i=1}^{n} p_i\right)}{\min\left(\sum_{i=1}^{n} p_i\right)}$$

over the time interval of interest, rather than as the greatest amplitude of any particular cycle in the interval.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 July; accepted 5 September 2008.
Published online 19 October 2008.

1. Winfree, A. T. *The Geometry of Biological Time* (Springer, 1980).
2. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Springer, 1984).
3. Izhikevich, E. M. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting* (MIT Press, 2007).
4. Blasius, B., Huppert, A. & Stone, L. Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature* **399**, 354–359 (1999).
5. Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Spigler, R. The Kuramoto model: A simple paradigm for synchronization phenomena. *Rev. Mod. Phys.* **77**, 137–185 (2005).
6. Watts, D. J. *Small Worlds* (Princeton Univ. Press, 1999).
7. Barahona, M. & Pecora, L. M. Synchronization in small-world systems. *Phys. Rev. Lett.* **89**, 054101 (2002).
8. Hong, H., Choi, M. Y. & Kim, B. J. Synchronization on small-world networks. *Phys. Rev. E* **65**, 026139 (2002).
9. Nishikawa, T., Motter, A. E., Lai, Y. C. & Hoppensteadt, F. C. Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize? *Phys. Rev. Lett.* **91**, 014101 (2003).
10. Hassell, M. P. & May, R. M. Aggregation of predators and insect parasites and its effect on stability. *J. Anim. Ecol.* **43**, 567–594 (1974).
11. Crowley, P. H. Dispersal and the stability of predator-prey interactions. *Am. Nat.* **118**, 673–701 (1981).
12. Comins, H. N. & Hassell, M. P. Persistence of multispecies host-parasitoid interactions in spatially distributed models with local dispersal. *J. Theor. Biol.* **183**, 19–28 (1996).
13. Ellner, S. P. *et al.* Habitat structure and population persistence in an experimental community. *Nature* **412**, 538–543 (2001).
14. Murdoch, W. W., Briggs, C. J. & Nisbet, R. M. *Consumer-Resource Dynamics* (Princeton Univ. Press, 2003).
15. Matthews, P. C., Mirollo, R. E. & Strogatz, S. H. Dynamics of a large system of coupled nonlinear oscillators. *Physica D* **52**, 293–331 (1991).
16. Crooks, K. R. & Sanjayan, M. A. *Connectivity Conservation* (Cambridge Univ. Press, 2006).
17. Jansen, V. A. A. The dynamics of two diffusively coupled predator-prey populations. *Theor. Popul. Biol.* **59**, 119–131 (2001).
18. Hudgens, B. R. & Haddad, N. M. Predicting which species will benefit from corridors in fragmented landscapes from population growth models. *Am. Nat.* **161**, 808–820 (2003).
19. Holyoak, M. Habitat patch arrangement and metapopulation persistence of predators and prey. *Am. Nat.* **156**, 378–389 (2000).
20. Urban, D. & Keitt, T. Landscape connectivity: a graph-theoretic perspective. *Ecology* **82**, 1205–1218 (2001).
21. Tewksbury, J. J. *et al.* Corridors affect plants, animals, and their interactions in fragmented landscapes. *Proc. Natl Acad. Sci. USA* **99**, 12923–12926 (2002).
22. Fortuna, M. A., Gómez-Rodriguez, C. & Bascompte, J. Spatial network structure and amphibian persistence in stochastic environments. *Proc. R. Soc. Lond. B* **273**, 1429–1434 (2006).
23. Hanski, I. & Ovaskainen, O. The metapopulation capacity of a fragmented landscape. *Nature* **404**, 755–758 (2000).
24. McIntire, E. J. B., Schultz, C. B. & Crone, E. E. Designing a network for butterfly habitat restoration: where individuals, populations and landscapes interact. *J. Appl. Ecol.* **44**, 725–736 (2007).
25. Hastings, A. Transient dynamics and persistence of ecological systems. *Ecol. Lett.* **4**, 215–220 (2001).
26. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
27. Rosenzweig, M. L. & MacArthur, R. H. Graphical representation and stability conditions of predator-prey interactions. *Am. Nat.* **97**, 209–223 (1963).
28. Hilty, J. A., Lidicker, W. Z. & Merenlender, A. M. *Corridor Ecology: The Science and Practice of Linking Landscapes for Biodiversity Conservation* (Island Press, 2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Holyoak for comments on an earlier version of the manuscript and D. Wysham for discussions. M.D.H. was funded by a Quantitative Environmental and Integrative Biology grant to A.H. and M. Holyoak from the National Science Foundation.

Author Contributions M.D.H. wrote custom software, ran simulations and analyzed data. M.D.H. and A.H. designed the study and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.D.H. (mdholland@ucdavis.edu).

METHODS

Numerical integration. Numerical integration was carried out using the backward-differentiation formula method in CVODE (ref. 29). Prey initial conditions were i.i.d. with $\log_{10}(h_i(0))$ uniformly distributed on the interval $(-5, 1 + \log_{10}\hat{h})$ and predator initial conditions were i.i.d. with $\log_{10}(p_i(0))$ uniformly distributed on the interval $(-5, 1 + \log_{10}\hat{p})$, where $\hat{h} = \eta/(\phi - \eta)$ and $\hat{p} = (1 + \hat{h})(1 - \theta\hat{h})$.

Dispersal networks. We used three types of dispersal networks (Fig. 1), all with an average number of connections per vertex (degree) of four. A ring lattice of degree four (Fig. 1a) is obtained by placing n points on a ring and connecting each to its four nearest neighbours. A rewired ring lattice (Fig. 1b) is obtained by deleting m randomly chosen edges from a ring lattice and replacing them at random. Because our networks are small, we chose random edges from the set of edges not included in the original lattice. We generated random graphs of degree four (Fig. 1c) by choosing $2n$ edges at random from the $n(n-1)/2$ possible undirected edges. All networks were connected (every vertex can be reached by traversing one or more edges from every other vertex).

Pseudorandom number generation. Pseudorandom numbers were generated using the Mersenne twister algorithm.

Cluster identification. Dynamics in pairs of patches were compared by calculating the linear correlation coefficient, ρ_{ij}, of prey time series in all pairs of patches. A pair of patches was taken to have identical dynamics, and thus assigned to the same cluster, if $\rho_{ij} > 0.999$.

Non-dimensionalization. The system of equations (1) and (2) was obtained by non-dimensionalizing a spatially explicit version of the Rosenzweig–MacArthur[27] equations

$$\frac{\mathrm{d}H_i}{\mathrm{d}\tau} = H_i\left(1 - \frac{H_i}{K}\right) - \frac{aP_iH_i}{b+H_i} + D_H\sum_{j=1}^{n} A_{ij}H_j$$

$$\frac{\mathrm{d}P_i}{\mathrm{d}\tau} = \frac{caP_iH_i}{b+H_i} - mP_i + D_P\sum_{j=1}^{n} A_{ij}P_j$$

by making the substitutions $h_i = H_i/b$, $p_i = aP_i/rb$ and $t = r\tau$ and defining the parameters $\theta = b/K$, $\phi = ca/r$, $\eta = m/r$, $d_h = D_H/r$ and $d_p = D_P/r$. The matrix $A = M - E$ represents the allowed dispersal transitions, where M is the adjacency matrix of the dispersal network, meaning that $M_{ij} = 1$ if individuals are allowed to move from patch j to patch i. E represents emigration, and is thus a diagonal matrix with entries $E_{ii} = \sum_{j=1}^{n} M_{ji}$.

Phase analysis of transient duration. Phase dynamics of solutions to the predator–prey model in equations (1) and (2) were reconstructed by treating the projections of the solution into the planes $h_i \times p_i$ as phase plane trajectories of limit cycle oscillators. The time between subsequent crossings from above of the line segment connecting the nontrivial equilibrium at $(\hat{h}, \hat{p})$ to the origin was taken to represent one cycle, or 2π radians of phase evolution. We defined $Z_i = f(T_i) - \overline{f(T_{i,\alpha})}$, where $T_i = T_{i,1}, T_{i,2}, \ldots, T_{i,j}, \ldots, T_{i,N}$ is the time series of N cycle periods in the ith patch, f is a low-pass filter and $T_{i,\alpha}$ is the series of periods in the ith patch for the last αN observations of T_i. Thus, the elements Z_i should be close to zero when asymptotic dynamics have been reached. In practice, it is necessary to choose a positive threshold, ε, and declare that asymptotic dynamics have been reached by time k if $|Z_i| < \varepsilon$ for all $j > k$. Errors can result when the asymptotic dynamics of one or more patches are poorly approximated by a simple limit cycle, but if these deviations are themselves periodic at high frequencies, they will be removed by the filter f. Most commonly, the one or two lowest amplitude patches will have sufficiently slow variation in period to cause errors. Therefore, in addition to filtering the T_i to obtain Z_i, we estimated the transient duration as the time until at least eight of ten patches had reached the stage of asymptotic behaviour by the above criterion, with $\varepsilon = 0.1$ and $\alpha = 0.167$.

The filter f was chosen to be a (Blackman) windowed sinc filter[30] with 151 points and a cut-off frequency of 0.05 per sample. This filter has a transition band from 0.04 (0-dB amplitude gain) to 0.07 per sample (−74-dB amplitude gain). In practice, this means that large-period deviations occurring on timescales of 14 or more cycles may remain in the Z_i, causing us to incorrectly conclude that asymptotic dynamics are never reached. The solution shown in Supplementary Figs 1i, 2i, 3i is an example of such a case.

29. Cohen, S. D. & Hindmarsh, A. C. CVODE, a stiff/nonstiff ODE solver in C. *Computers Phys.* **10**, 138–143 (1996).
30. Smith, S. W. *The Scientist and Engineer's Guide to Digital Signal Processing* (California Technical Publishing, 1997).

LETTERS

Coordinate control of synaptic-layer specificity and rhodopsins in photoreceptor neurons

Marta Morey[1]*, Susan K. Yee[1]*, Tory Herman[1]†, Aljoscha Nern[1], Enrique Blanco[2] & S. Lawrence Zipursky[1]

How neurons make specific synaptic connections is a central question in neurobiology. The targeting of the *Drosophila* R7 and R8 photoreceptor axons to different synaptic layers in the brain provides a model with which to explore the genetic programs regulating target specificity. In principle this can be accomplished by cell-type-specific molecules mediating the recognition between synaptic partners[1]. Alternatively, specificity could also be achieved through cell-type-specific repression of particular targeting molecules. Here we show that a key step in the targeting of the R7 neuron is the active repression of the R8 targeting program. Repression is dependent on NF-YC, a subunit of the NF-Y (nuclear factor Y) transcription factor[2]. In the absence of NF-YC, R7 axons terminate in the same layer as R8 axons. Genetic experiments indicate that this is due solely to the derepression of the R8-specific transcription factor Senseless[3] (Sens) late in R7 differentiation. Sens is sufficient to control R8 targeting specificity and we demonstrate that Sens directly binds to an evolutionarily conserved DNA sequence upstream of the start of transcription of an R8-specific cell-surface protein, Capricious (Caps) that regulates R8 target specificity. We show that R7 targeting requires the R7-specific transcription factor Prospero[4,5] (Pros) in parallel to repression of the R8 targeting pathway by NF-YC. Previous studies demonstrated that Sens[6,7] and Pros[8] directly regulate the expression of specific rhodopsins in R8 and R7. We propose that the use of the same transcription factors to promote the cell-type-specific expression of sensory receptors and cell-surface proteins regulating synaptic target specificity provides a simple and general mechanism for ensuring that transmission of sensory information is processed by the appropriate specialized neural circuits.

The compound eye comprises about 750 simple eyes (ommatidia), each containing a cluster of eight photoreceptor neurons (R1–R8). These neurons form synaptic connections in two regions of the optic lobe, the lamina and the medulla (Fig. 1a). The R1–R6 neurons innervate the lamina; the R7 and R8 neurons form connections in the M6 and M3 medulla layers, respectively. Genetic studies have led to the identification of cell-surface proteins regulating R7 and R8 target specificity[9–18]. Notably, mis-targeting mutant R7 neurons terminate selectively in M3 (refs 9–14), the layer in which wild-type R8 axons terminate, suggesting a close relationship between the genetic programs controlling R7 and R8 target specificity. Here we describe transcriptional regulatory pathways that control the differential targeting specificity of these neurons.

In a screen for R7 targeting mutants[10], we identified a strong loss of function mutation in the *NF-YC* gene (Supplementary Fig. 1), which encodes a subunit of NF-Y, an evolutionarily conserved heterotrimeric transcription factor. Although NF-Y function has not been studied extensively in the fly[19,20], it has been shown to act as both an activator[21] and a repressor[22] in other organisms. The targeting of visual-system neurons was assessed in mosaic animals[9] to generate large patches of mutant retinal tissue early in development (Fig. 1b, b′). About 75% of *NF-YC* mutant R7 axons ($n = 892$ of 1,148)

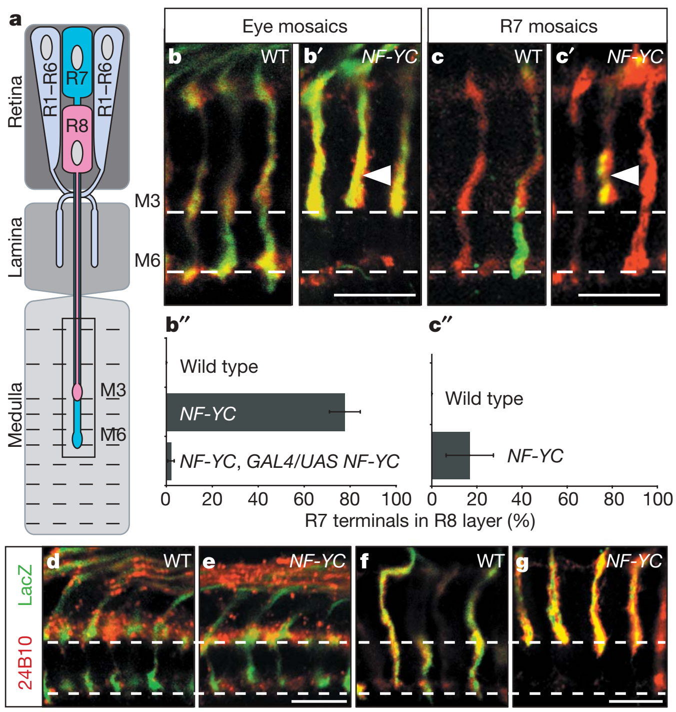


Figure 1 | *NF-YC* mutant R7 axons mis-target to the M3 layer. **a**, Schematic diagram of photoreceptor neurons (R cells) in the *Drosophila* visual system. The box indicates the area shown in **b**, **b′**, **c**, **c′** and **d**–**g**. **b**–**c″**, Targeting of *NF-YC* mutant R7s (green) in mosaic animals (**b**–**b″**, *ey3.5FLP*; **c**–**c″**, *GMRFLP*). Monoclonal antibody 24B10 (red) labelled all R7 and R8 axons. Arrowheads, mutant R7s terminating in M3. Scale bar, 10 μm. Error bars indicate s.d. **b**–**b″**, Large patches of control (wild type (WT), **b**) and mutant (*NF-YC*, **b′**) R cells. WT R7s target exclusively to M6 (11 brains, 895 R7 neurons; **b**, **b″**), whereas *NF-YC* mutant axons target to M3 (77.7 ± 6.7%, 8 brains, 1,148 R7s; **b′**, **b″**). The mis-targeting phenotype was rescued by the expression of *NF-YC* cDNA (2.0 ± 1.4%, 6 brains, 1,076 R7s; **b″**). **c**–**c″**, All WT R7s targeted to M6 (9 brains, 812 R7s; **c**, **c″**), whereas isolated mutant R7s (16.8 ± 10.5%, 10 brains, 807 R7s; **c′**, **c″**) targeted to M3. **d**–**g**, Targeting of WT and *NF-YC* mutant R7s at different developmental stages (green). At 30% APF, both WT (**d**) and mutant (**e**) R7s target correctly to their temporary layer. At 70% APF, WT R7s (**f**) target to the M6 layer. Mutant R7s (**g**) terminate in M3. Scale bar, 10 μm.

[1]Department of Biological Chemistry, Howard Hughes Medical Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA. [2]Departament de Genètica and Institut de Biomedicina de la Universitat de Barcelona (IBUB), Universitat de Barcelona, 08028 Barcelona, Catalunya, Spain. †Present address: Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA.
*These authors contributed equally to this work.

terminated in M3, the same layer as wild-type R8 axons (Fig. 1b′, b″). This phenotype was fully rescued by an *NF-YC* complementary DNA (Fig. 1b″). In contrast with the marked effect of *NF-YC* mutations on R7, targeting of R8 to the M3 layer and targeting of R1–R6 to the lamina were unaffected (Supplementary Fig. 2).

To assess whether *NF-YC* is required in a cell-autonomous fashion in R7 neurons, we generated mosaic flies in which a fraction of R7 neurons was rendered mutant and labelled with green fluorescent protein (GFP), whereas the remaining R7 neurons and all the R8 neurons were wild-type and unlabelled (Fig. 1c, c′). We observed that about 17% of the mutant R7 neurons ($n = 144$ of 807) mis-targeted to M3 (Fig. 1c′, c″). The decrease in penetrance of the phenotype, in comparison with mutant R7 neurons generated by mitotic recombination induced earlier in the eye primordium (compare Fig. 1b″ with Fig. 1c″), probably reflects perdurance of NF-YC protein present in precursor cells. *NF-YC* is therefore required autonomously for R7 targeting but not for the targeting of other classes of photoreceptor neurons. As NF-YC is expressed in all R cells (Supplementary Fig. 3), NF-YC must function in combination with other factors or signals selectively activating NF-YC function in R7.

Given that NF-YC is part of a transcription factor complex[20] and is expressed in the nucleus of R7 neurons, it is likely that the change in targeting specificity reflects a change in gene expression. Wild-type R7 neurons initially target to the temporary R7 layer in the medulla and then, during mid-pupal development, extend to their final target layer[11]. Targeting of *NF-YC* mutant R7 neurons to the temporary layer is indistinguishable from the wild type (Fig. 1d, e). Extension to the final target layer at 70% after puparium formation (APF) is frequently disrupted, with many R7 neurons terminating in the layer within which R8 terminates (Fig. 1f, g). Consistent with this finding was our observation that *NF-YC* mutant R7 neurons expressed all five early R7 markers tested in wild-type patterns (Table 1). We reasoned, then, that *NF-YC* might repress a subset of R8-specific genes in the R7 neuron that later in development control final target layer selection. Indeed, the R8-specific transcription factor Sens was expressed ectopically in *NF-YC* mutant R7 neurons (Table 1 and Fig. 2).

sens is a key regulator of R8 development[23]. In wild-type larval eye discs, Sens is expressed in two or three cells that have the potential to become R8 before becoming restricted to a single differentiating R8

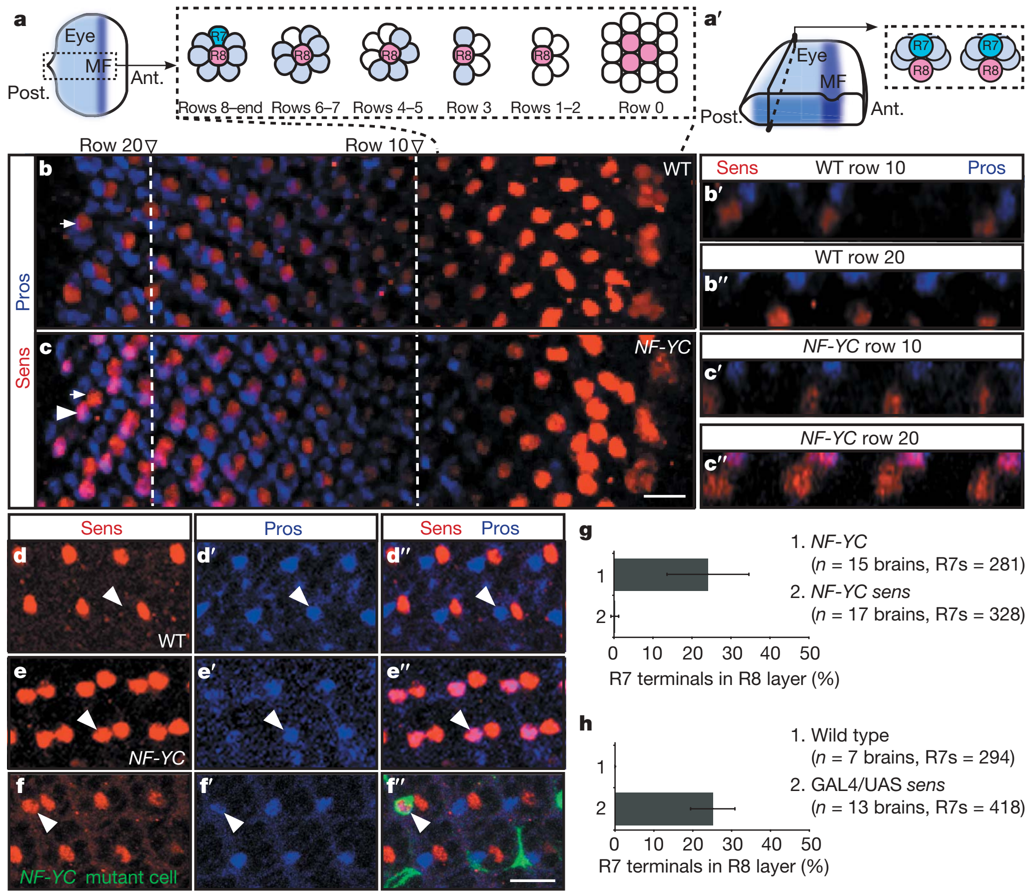


Figure 2 | Mis-expression of Sens in *NF-YC* mutant R7s is necessary and sufficient for the R7 mis-targeting phenotype. **a**, **a′**, Diagrams of R-cell recruitment to the ommatidium in the eye disc. **a**, Eye disc viewed from above the epithelium. R-cell clusters form as a wave of differentiation moves from posterior (post.) to anterior (ant.) with specific R cells joining the cluster at different times (that is, correlating with the distance from row 0, morphogenetic furrow (MF)). Red, R8 nuclei; blue, R7 nuclei; white, undifferentiated cells; light blue, R1–R6. During development, clusters rotate by 90°. **a′**, Cross-section along the apicobasal axis of the disc. **b**–**c″**, WT (**b**–**b″**) and *NF-YC* mutant (**c**–**c″**) eye discs stained for Sens (red) and Pros (blue). **b**–**b″**, In WT, Sens is expressed in two or three precursor cells before becoming restricted to a single R8 (arrow, **b**). Pros expression in R7 begins as R7 differentiation commences (about row 8). Pros is expressed at later stages in non-neuronal cells. Cross-sections are shown in **b′** and **b″**. **c**–**c″**, *NF-YC* mutant R7s express Pros like WT R7s. Sens is detected in mutant R7s 10–12 rows after R7 differentiation (arrowhead, **c**). R8s express Sens in a WT pattern (arrow). Co-expression of Sens and Pros in mutant R7s is seen at row 20 (**c″**), but not at row 10 (**c′**). Scale bar, 10 μm. **d**–**f″**, WT (**d**–**d″**), mutant (**e**–**e″**), and mosaic (**f**–**f″**) retinas at 40% APF stained for Sens (red) and Pros (blue). Arrowheads, R7s. **d**–**d″**, WT R8s express Sens, whereas WT R7s express Pros. **e**–**e″**, *NF-YC* mutant R7s also express Sens (**e″**). **f**–**f″**, In mosaics, single *NF-YC* mutant R7s (arrowhead, **f″**) mis-express Sens. Scale bar, 10 μm. **g**, **h**, Sens is necessary (**g**) and sufficient (**h**) for R7 targeting to M3 in *NF-YC* mutants. Error bars indicate s.d. **g**, Mis-targeting of *NF-YC sens* double-mutant R7s (0.2 ± 1.0%) and *NF-YC* control (24.1 ± 10.5%). **h**, Mis-targeting of WT R7s mis-expressing Sens (25.1 ± 5.7%).

Table 1 | Marker expression in *NF-YC* mutant eye discs

Markers	Name	Wild-type R8	*NF-YC* R8	Wild-type R7	*NF-YC* R7
R7 markers	H214 (Klingon)*	–	–	+	+
	PM181-GAL4 (Sevenless)†	–	–	+	+
	Prospero‡	–	–	+	+
	Deadpan-LacZ†	–	–	+	+
R7 & R8 markers	Runt‡	+	+	+	+
R8 markers	Atonal‡	+	+	–	–
	Boss‡	+	+	–	–
	BBO2*	+	+	–	–
	RO-156*	+	+	–	–
	109-68 (Scabrous)*	+	+	–	–
	Senseless‡	+	+	–	+

Markers were analysed by using enhancer trap lines (*), promoter fusion (†) or antibodies (‡).

neuron[23] (Fig. 2a–b″). Sens remains expressed in R8 into the adult[6,7] (see Supplementary Fig. 6b). It is required at a very early stage of eye development to regulate R8 specification[23] and, much later during pupal development and in the adult, to regulate the transcription of R8-specific rhodopsins directly[7]. In *NF-YC* mutant larval eye discs, Sens expression in R8 begins before overt R8 differentiation as in the wild type. By contrast, Sens mis-expression in mutant R7 neurons (Fig. 2c) was first observed 15–20 h after the onset of differentiation as assessed by the expression of the R7-specific marker *pros* (Table 1, and compare Fig. 2c′ with Fig. 2c″). Expression of Sens in mutant R7 neurons persists throughout pupal development (Fig. 2d–e″) and into the adult (see Supplementary Fig. 6d) and is cell-autonomous (Fig. 2f–f″). As Sens mis-expression occurs after the onset of R7 differentiation and *NF-YC* mutant R7 neurons mis-target to the M3 layer during the late phase of R7 targeting, *sens* may promote an R8 targeting program that is distinct from the role of *sens* in cell fate earlier in development.

If upregulation of Sens in *NF-YC* mutant R7 neurons is responsible for targeting to the M3 layer, removal of Sens from *NF-YC* mutant cells should suppress the targeting defect. To test this, we induced mitotic recombination on two different chromosomes (namely chromosomes X and 3) to generate R7 neurons that were simultaneously mutant for both *NF-YC* and *sens*, and we assessed their targeting in an otherwise wild-type background (Supplementary Fig. 4a). Removing *sens* from *NF-YC* mutant R7 neurons completely suppressed the mis-targeting phenotype (Fig. 2g). Thus, during wild-type development the *NF-YC* mediated repression of *sens* in R7 is necessary to prevent inappropriate targeting to M3.

To test whether Sens is sufficient to implement an R8 targeting program, we mis-expressed *sens* in R7 neurons (Supplementary Fig. 4b; see Methods). Under these conditions about 25% ($n = 104$ of 418) of the R7 neurons were redirected to the M3 layer, thus phenocopying *NF-YC* loss-of-function mutations (Fig. 2h). Additional experiments using the method[7] in which Sens was provided conditionally early in development to promote R8 cell fate, but removed later, support the view that Sens functions at later stages of R8 development to promote targeting (Supplementary Fig. 5). Taken together, these data raise the possibility that Sens could directly control the expression of cell-surface proteins regulating R8 target specificity.

Caps is the only cell-surface molecule that has been shown to be both specifically expressed in the R8 neuron and required for R8 targeting[18] and it is therefore an excellent candidate for direct regulation by Sens. Indeed, like Sens, Caps is expressed ectopically in R7 in *NF-YC* mutants (Fig. 3a–f). Expression of Caps, as detected with an enhancer trap, is specifically activated in *NF-YC* mutant R7 neurons about 9 h after the onset of Sens expression. Furthermore, a previous study[18] showed that ectopic expression of Caps in R7 respecified their connections to the R8 layer. Both *NF-YC* mutant R7 neurons and R7 neurons mis-expressing Caps initially target correctly but then select the inappropriate M3 layer during mid-pupal development. Taken together, these observations indicate that *caps* could be a downstream target of Sens.

Examination of the DNA sequences 1 kilobase upstream of *caps* and within the first large intron led to the identification of four and three putative Sens-binding sites, respectively. We identified an

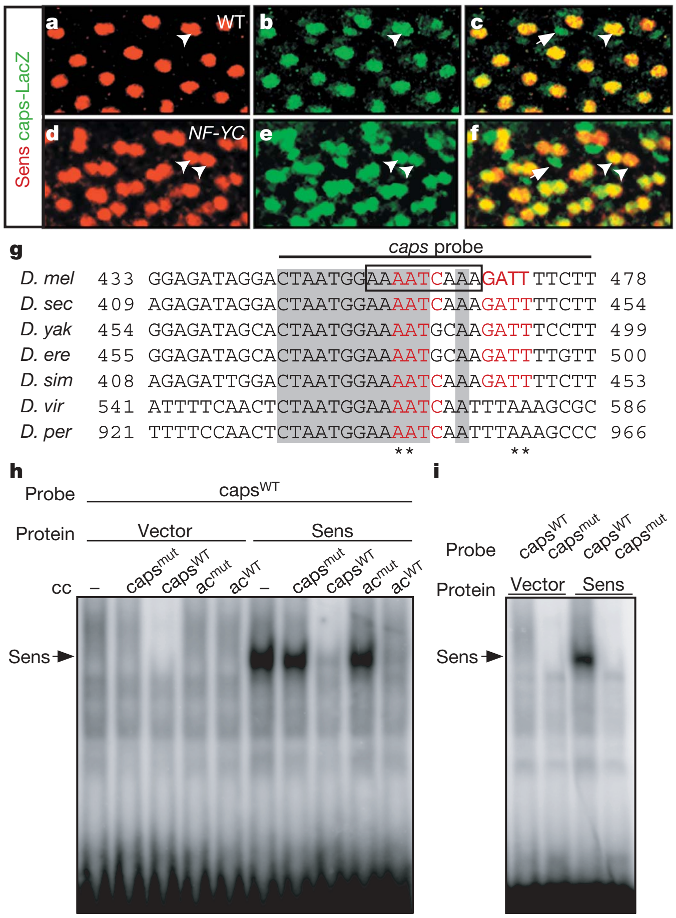


Figure 3 | Sens regulates *caps* expression. **a–f**, WT (**a–c**) and *NF-YC* mutant (**d–f**) eye discs carrying *caps-lacZ*, stained for Sens (red) and LacZ (green). **a–c**, In WT, R8s express LacZ (**c**, arrowhead). **d–f**, In *NF-YC* mutant, LacZ is also detected in R7s (**f**, arrowheads). Arrows indicate LacZ expression in non-neuronal cone cells both in WT (**c**) and in *NF-YC* mutant tissue (**f**). Original magnification, ×40 (**a–f**). **g**, Alignment showing conservation (grey) of Sens-binding site (box) in *caps*. In closely related species there is another predicted Sens-binding site (core sequence, red). Line, *caps* probe in **h** and **i**. Asterisk, AA→CC nucleotide change in the mutant probe. *D.mel*, *Drosophila melanogaster*; *D.sec*, *Drosophila sechellia*; *D.yak*, *Drosophila yakuba*; *D.ere*, *Drosophila erecta*; *D.sim*, *Drosophila simulans*; *D.vir*, *Drosophila virilis*; *D.per*, *Drosophila persimilis*. **h**, **i**, Sens binds to predicted Sens sites on *caps* probe. **h**, Gel-shift assay with WT *caps* probe and either 'vector' control or Sens protein. cc, 50× unlabelled competitor probes added as indicated. $caps^{WT}$, *caps* probe; ac^{WT}, *achaete* probe[29] (positive control); $caps^{mut}$ and ac^{mut}, *caps* and *ac* probes with Sens-binding sites mutated. **i**, Gel-shift assay with WT or mutant *caps* probes.

evolutionarily conserved Sens-binding site 500 base pairs upstream of the *caps* transcriptional start site (Fig. 3g; see Supplementary Methods). Sens protein binds specifically to this site in gel-shift assays, making it likely that *caps* is a direct target of Sens (Fig. 3h, i). However, Sens must regulate R8 target specificity by controlling the expression of other genes in addition to *caps*, because loss of *caps* does not suppress the *NF-YC* mutant phenotype (*NF-YC caps* 20.3 ± 11.0%, 23 brains, 441 R7 neurons; *NF-YC* 24.1 ± 10.5%, 15 brains, 281 R7 neurons; means ± s.d.). This is consistent with the finding that loss of *caps*, in an otherwise wild-type background, results in targeting defects in about 50% of the R8 neurons[18]. Together, these data suggest that Sens directly regulates the expression of Caps, a cell-surface protein controlling R8 target specificity, and must also regulate the expression of other genes involved in this process.

Specific repression of *sens* in R7 neurons could arise through interactions between NF-YC and the R7-specific transcription factor Pros. Like NF-YC, Pros is also required for R7 target specificity[24,25]. It is expressed in R7 from an early stage of its development through to the adult[8,24] in a similar fashion to Sens expression in R8. About 20% of the *pros*-null mutant R7 neurons terminate in M3 (Fig. 4a). Two lines of evidence support the view that Pros works in parallel with NF-YC: first, the loss of *pros* in R7 neurons does not lead to ectopic expression of Sens (Fig. 4b, c), and second, the frequency of mis-targeting R7 axons in single *pros*-null mutant cells is markedly increased by removing *NF-YC* (Fig. 4a; see Methods). Thus, Pros could either promote R7 targeting directly or, like NF-YC, act to repress an R8 targeting program, or both.

Thus, R7 targeting requires NF-YC and, in parallel, Pros, whereas R8 targeting relies on Sens-dependent regulation of *caps* and other genes. Mutations in many other genes required for R7 targeting cause R7 neurons to mis-target to the M3 layer specifically rather than terminating promiscuously in the medulla. This underscores a tight inter-relationship between the mechanisms regulating targeting to these two layers. On the basis of the strong M3 mis-targeting phenotype of *NF-YC* mutant R7 neurons and complete suppression of the phenotype by the removal of *sens*, a key mechanism regulating R7 layer specificity is repression of an R8 targeting program. More generally, repression of inappropriate pathways may promote differential targeting in closely related neurons[26].

The roles of Pros and Sens in target layer selection are analogous to their function in controlling the expression of R7-specific and R8-specific rhodopsins. R7 and R8 neurons express different rhodopsins and hence detect different wavelengths of light. In R8, Sens directly represses the transcription of R7 rhodopsins and directly activates the transcription of an R8 rhodopsin[7]. In the R7 neuron, Pros binds to an upstream regulatory sequence in the R8 rhodopsin genes and represses their expression[7,8]. *NF-YC* mutant R7 neurons no longer express R7 rhodopsins, and all express R8 rhodopsins (Supplementary Fig. 6). This is consistent with the finding that *NF-YC* mutant R7 neurons in adults express Sens but no longer express Pros. Thus, transcription of both R8-specific rhodopsins and, as we have shown here, an R8-specific targeting protein Caps is directly regulated by Sens (Fig. 4d).

These observations suggest a simple solution to the mechanisms by which sensory neurons connect to the neural circuits specialized for the reception of different sensory stimuli (for example, different wavelengths of light or different odours). Although the molecular basis of this coupling is understood in considerable detail for vertebrate olfactory neurons, in which odorant receptors have a direct function in controlling target specificity[27], little is known about the coupling in other sensory systems. Coupling is likely to be regulated in a different fashion in other neurons, because even in the fly olfactory system, for example, targeting is independent of sensory receptor expression[28]. On the basis of our studies on Sens we propose that the same transcription factors directly control both rhodopsin expression and the cell-surface proteins that control target layer specificity. More generally, we speculate that in many sensory neurons a common set of transcription factors may directly control, and thereby coordinate, the expression of cell-surface proteins regulating target specificity and the receptors detecting specific sensory stimuli.

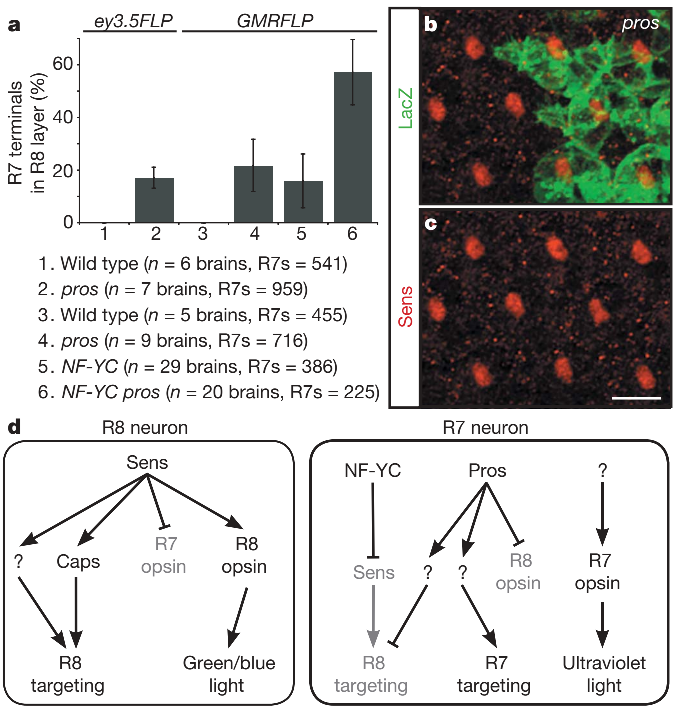


Figure 4 | Pros and NF-YC regulate R7 targeting in parallel pathways. **a**, *pros* and *NF-YC pros* R7 phenotypes in large (*ey3.5FLP*) and isolated (*GMRFLP*) patches of mutant tissue. WT R7s target to M6 (1, 3), whereas *pros* R7s mis-target to M3 (2, 17.1 ± 4%; 4, 21.8 ± 9.9%). *NF-YC pros* R7s have a more severe mis-targeting phenotype (6, 57.2 ± 12.4%; *NFYC* control, 5, 15.9 ± 10.2%). Error bars indicate s.d. **b**, **c**, A 40% APF retina with large patches of *pros* mutant cells (green) stained for Sens (red). R8s express Sens; *pros* mutant R7s do not mis-express Sens. Scale bar, 10 μm. **d**, Summary of the genetic control of layer specificity and rhodopsin expression in R8 and R7 neurons.

METHODS SUMMARY

To generate *NF-YC* or *pros* mutant R7s, we used the MARCM (mosaic analysis with a repressible cell marker) technique as described previously[10]. To generate R7 mosaics (that is, mutant R7 neurons surrounded by wild-type cells) *GMR-FLP* was used as a source of recombinase. For larger patches of mutant tissue in the eye, including R7 neurons (eye mosaics), the *ey3.5-FLP* construct was used. A similar version of the MARCM technique with *GMR-FLP* was used to generate double-mutant R7 neurons. Mutant R7 neurons were marked only when flipping occurred on both chromosomes X and 3. *GMR-FLP* was used with MARCM to generate R7 neurons expressing Sens (*UAS-sens*), in addition to the inducible marker. For assessing expression of R7 and R8 markers, hemizygous *NF-YC*[40] mutant third-instar eye discs and late pupae were analysed.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 June; accepted 10 September 2008.
Published online 2 November 2008; corrected 11 December 2008 (details online).

1. Yamagata, M., Weiner, J. A. & Sanes, J. R. Sidekicks: synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell* **110**, 649–660 (2002).
2. Mantovani, R. The molecular biology of the CCAAT-binding factor NF-Y. *Gene* **239**, 15–27 (1999).
3. Nolo, R., Abbott, L. A. & Bellen, H. J. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila*. *Cell* **102**, 349–362 (2000).
4. Doe, C. Q., Chu-LaGraff, Q., Wright, D. M. & Scott, M. P. The *prospero* gene specifies cell fates in the *Drosophila* central nervous system. *Cell* **65**, 451–464 (1991).
5. Vaessin, H. *et al. prospero* is expressed in neuronal precursors and encodes a nuclear protein that is involved in the control of axonal outgrowth in *Drosophila*. *Cell* **67**, 941–953 (1991).

6. Domingos, P. M. *et al.* Regulation of R7 and R8 differentiation by the *spalt* genes. *Dev. Biol.* **273,** 121–133 (2004).
7. Xie, B., Charlton-Perkins, M., McDonald, E., Gebelein, B. & Cook, T. Senseless functions as a molecular switch for color photoreceptor differentiation in *Drosophila. Development* **134,** 4243–4253 (2007).
8. Cook, T., Pichaud, F., Sonneville, R., Papatsenko, D. & Desplan, C. Distinction between color photoreceptor cell fates is controlled by Prospero in *Drosophila. Dev. Cell* **4,** 853–864 (2003).
9. Lee, C. H., Herman, T., Clandinin, T. R., Lee, R. & Zipursky, S. L. N-cadherin regulates target specificity in the *Drosophila* visual system. *Neuron* **30,** 437–450 (2001).
10. Nern, A. *et al.* An isoform-specific allele of *Drosophila N-cadherin* disrupts a late step of R7 targeting. *Proc. Natl Acad. Sci. USA* **102,** 12944–12949 (2005).
11. Ting, C. Y. *et al. Drosophila* N-cadherin functions in the first stage of the two-stage layer-selection process of R7 photoreceptor afferents. *Development* **132,** 953–963 (2005).
12. Clandinin, T. R. *et al. Drosophila* LAR regulates R1–R6 and R7 target specificity in the visual system. *Neuron* **32,** 237–248 (2001).
13. Maurel-Zaffran, C., Suzuki, T., Gahmon, G., Treisman, J. E. & Dickson, B. J. Cell-autonomous and -nonautonomous functions of LAR in R7 photoreceptor axon targeting. *Neuron* **32,** 225–235 (2001).
14. Newsome, T. P., Asling, B. & Dickson, B. J. Analysis of *Drosophila* photoreceptor axon guidance in eye-specific mosaics. *Development* **127,** 851–860 (2000).
15. Lee, R. C. *et al.* The protocadherin Flamingo is required for axon target selection in the *Drosophila* visual system. *Nature Neurosci.* **6,** 557–563 (2003).
16. Senti, K. A. *et al.* Flamingo regulates R8 axon–axon and axon–target interactions in the *Drosophila* visual system. *Curr. Biol.* **13,** 828–832 (2003).
17. Tomasi, T., Hakeda-Suzuki, S., Ohler, S., Schleiffer, A. & Suzuki, T. The transmembrane protein Golden goal regulates R8 photoreceptor axon–axon and axon–target interactions. *Neuron* **57,** 691–704 (2008).
18. Shinza-Kameda, M., Takasu, E., Sakurai, K., Hayashi, S. & Nose, A. Regulation of layer-specific targeting by reciprocal expression of a cell adhesion molecule, Capricious. *Neuron* **49,** 205–213 (2006).
19. Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111,** 687–701 (2002).
20. Yoshioka, Y. *et al.* Complex interference in the eye developmental pathway by *Drosophila* NF-YA. *Genesis* **45,** 21–31 (2007).
21. Frontini, M. *et al.* NF-Y recruitment of TFIID, multiple interactions with histone fold TAF_{II}s. *J. Biol. Chem.* **277,** 5841–5848 (2002).
22. Peng, Y. & Jahroudi, N. The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. *J. Biol. Chem.* **278,** 8385–8394 (2003).
23. Frankfort, B. J., Nolo, R., Zhang, Z., Bellen, H. & Mardon, G. *senseless* repression of *rough* is required for R8 photoreceptor differentiation in the developing *Drosophila* eye. *Neuron* **32,** 403–414 (2001).
24. Kauffmann, R. C., Li, S., Gallagher, P. A., Zhang, J. & Carthew, R. W. Ras1 signaling and transcriptional competence in the R7 cell of *Drosophila. Genes Dev.* **10,** 2167–2178 (1996).
25. Miller, A. C., Seymour, H., King, C. & Herman, T. G. Loss of *seven-up* from *Drosophila* R1/R6 photoreceptors reveals a stochastic fate choice that is normally biased by Notch. *Development* **135,** 707–715 (2008).
26. Kania, A., Johnson, R. L. & Jessell, T. M. Coordinate roles for LIM homeobox genes in directing the dorsoventral trajectory of motor axons in the vertebrate limb. *Cell* **102,** 161–173 (2000).
27. Mombaerts, P. *et al.* Visualizing an olfactory sensory map. *Cell* **87,** 675–686 (1996).
28. Dobritsa, A. A., van der Goes van Naters, W., Warr, C. G., Steinbrecht, R. A. & Carlson, J. R. Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. *Neuron* **37,** 827–841 (2003).
29. Jafar-Nejad, H. *et al.* Senseless acts as a binary switch during sensory organ precursor selection. *Genes Dev.* **17,** 2966–2978 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. N. Jan, H. J. Bellen, A. Nose, S. Britt, C. Desplan, T. Cook, U. Banerjee, the Developmental Studies Hybridoma Bank and Bloomington Stock Center for reagents; H. Schjerven, S. T. Smale and M. Carey for advice and assistance with gel-shift assays; and members of the Zipursky laboratory for comments on the manuscript. S.L.Z. is an investigator of the Howard Hughes Medical Institute. M.M. was supported by the Howard Hughes Medical Institute. S.K.Y. was supported by the Ruth L. Kirschstein National Research Service Award GM7185. T.H. was supported by the Burroughs Wellcome Fund for Biomedical Research. A.N. was supported by the European Molecular Biology Organization and the Human Frontiers Science Program. E.B. was supported by Juan de la Cierva postdoctoral contract and the grant BMC2006-07334 from the Ministerio de Educación y Ciencia (MEC), Spain.

Author Contributions M.M., S.K.Y. and S.L.Z. conceived the experiments and wrote the paper. M.M. and S.K.Y. performed all the experiments. The X-chromosome mosaic screen was designed and performed by T.H. and A.N. E.B. performed all computational sequence analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.L.Z. (lzipursky@mednet.ucla.edu).

METHODS

The stocks used to assess the targeting of *NF-YC* mutant R7 neurons in large clones were *ey3.5-FLP, tubP-GAL80, FRT19; act-GAL4, UAS-lacZ/CyO* and *NF-YC^{40}, FRT19/FM7, Kr-GFP*. To assess whether *NF-YC* was autonomously required in R7 neurons we used *GMR-FLP*, which induces mitotic recombination in R7, but not R8 cell precursors. The stocks used were *GMR-FLP, tubP-GAL80, FRT19; act-GAL4, UAS-syt-GFP/CyO* and *NF-YC^{40}, FRT19/FM7, Kr-GFP*. *FRT19/FRT19* was used as a control. Because *NF-YC^{40}* mutant animals die before eclosion, analysis of R7 and R8 markers was performed on hemizygous mutant *NF-YC^{40} FRT19/Y* and wild-type *FRT19* eye discs and pupal retinas. To assess the possibility that mis-expression of *sens* in R7 clones was autonomous we used the same stocks used to assess the *NF-YC* requirement in R7. The stocks used to generate flies with R7 neurons double mutant for *NF-YC* and *sens* or *NF-YC* and *caps* were *NF-YC^{40}, FRT19/FM7, Kr-GFP; +; sens^{E1}, FRT79/TM6B* or *NF-YC^{40}, FRT19/FM7, Kr-GFP; +; caps^{C18fs}, FRT79/TM6B* crossed to *GMRFLP, tubP-GAL80, FRT19; act-GAL4, UAS-syt-GFP/CyO; tubP-GAL80, FRT79/TM6B*. A similar scheme was used to generate single R7 neurons mutant for both *NF-YC* and *pros* (*FRT82B, pros$^{17.17}$*). As a control, to generate animals with R7 neurons mutant for *NF-YC* only, *NF-YC^{40}, FRT19/FM7, Kr-GFP; +; FRT79* or *FRT82B* flies were crossed to *GMRFLP, tubP-GAL80, FRT19; act-GAL4, UAS-syt-GFP/CyO; tubP-GAL80, FRT79* or *tubP-GAL80 FRT82B/TM6B*. The stocks used to mis-express *sens* in clones of wild-type R7 neurons were *FRT40/FRT40; UAS-sens^{C5}/TM6B* and *GMR-FLP, tubP-GAL80, FRT40/CyO; tub-GAL4, UAS-Nsyb-GFP/TM6B*. Similar results were obtained with independent insertions on chromosomes X (*UAS-sens^{C1}*) and 3 (*UAS-sens^{C6}*). We confirmed that wild-type R7 neurons mis-expressing *sens* also expressed the R7-specific marker Pros at 40% APF. As a control we generated clones of wild-type R7 neurons by crossing *FRT40/FRT40* to *GMR-FLP, tubP-GAL80, FRT40/CyO; tub-GAL4, UAS-Nsyb-GFP/TM6B*.

LETTERS

Temporal identity in axonal target layer recognition

Milan Petrovic[1] & Thomas Hummel[1]

The segregation of axon and dendrite projections into distinct synaptic layers is a fundamental principle of nervous system organization and the structural basis for information processing in the brain[1]. Layer-specific recognition molecules that allow projecting neurons to stabilize transient contacts and initiate synaptogenesis[2,3] have been identified. However, most of the neuronal cell-surface molecules critical for layer organization are expressed broadly in the developing nervous system[4,5], raising the question of how these so-called permissive adhesion molecules support synaptic specificity. Here we show that the temporal expression dynamics of the zinc-finger protein sequoia is the major determinant of *Drosophila* photoreceptor connectivity into distinct synaptic layers. Neighbouring R8 and R7 photoreceptors show consecutive peaks of elevated sequoia expression, which correspond to their sequential target-layer innervation. Loss of *sequoia* in R7 leads to a projection switch into the R8 recipient layer, whereas a prolonged expression in R8 induces a redirection of their axons into the R7 layer. The *sequoia*-induced axon targeting is mediated through the ubiquitously expressed Cadherin-N cell adhesion molecule. Our data support a model in which recognition specificity during synaptic layer formation is generated through a temporally restricted axonal competence to respond to broadly expressed adhesion molecules. Because developing neurons innervating the same target area often project in a distinct, birth-order-dependent sequence, temporal identity seems to contain crucial information in generating not only cell type diversity during neuronal division[6] but also connection diversity of projecting neurons.

In the *Drosophila* compound eye, each ommatidium contains eight photoreceptor neurons (R1–R8) that form synapses in distinct optic lobe ganglia, the peripheral lamina and the deeper medulla[7] (Fig. 1a). Axons of the colour-sensitive R8 and R7 cells project into the medulla and segregate into two out of ten synaptic layers, M3 and M6 respectively (Fig. 1a–c, f). In a mosaic screen for genes that control R8/R7 target layer selection we identified mutants of *sequoia* (*seq*[5] and *seq*[6]) with a frequent loss of R-cell innervation in M3 and M6 (100% of the analysed adult brains, $n > 40$ for each mutant; Supplementary Fig. 1c, d). Using the MARCM (mosaic analysis with a repressible cell marker) technique[8] for the selective labelling of homozygous mutant R8 and R7 cell axons, we found that more than 90% of *seq* mutant R7 cells terminate their axon projections in the outer M1–M3 layers ($n > 180$), correlating with an innervation gap in the M6 layer of the same medulla column (Fig. 1d, g). Similarly, *seq* mutant R8 axons frequently (87%, $n > 150$) stop above their prospective M3 target layer (Fig. 1g). The analysis of different R8/R7 genetic mosaics (Supplementary Fig. 2) showed that wild-type R8/R7 axons are not influenced by a *seq* mutant neighbour axon of the same medulla column, suggesting independent target recognition by these two R cells.

sequoia encodes a nuclear protein with two putative zinc-finger domains related to the DNA-binding domain of Tramtrack[9]. In the visual system, the four independently isolated *sequoia* alleles seq^{5}, seq^{6}, seq^{22} and seq^{SH1898} lead to indistinguishable R8/R7 axon-targeting phenotypes (Supplementary Fig. 1c–f). No expression could be

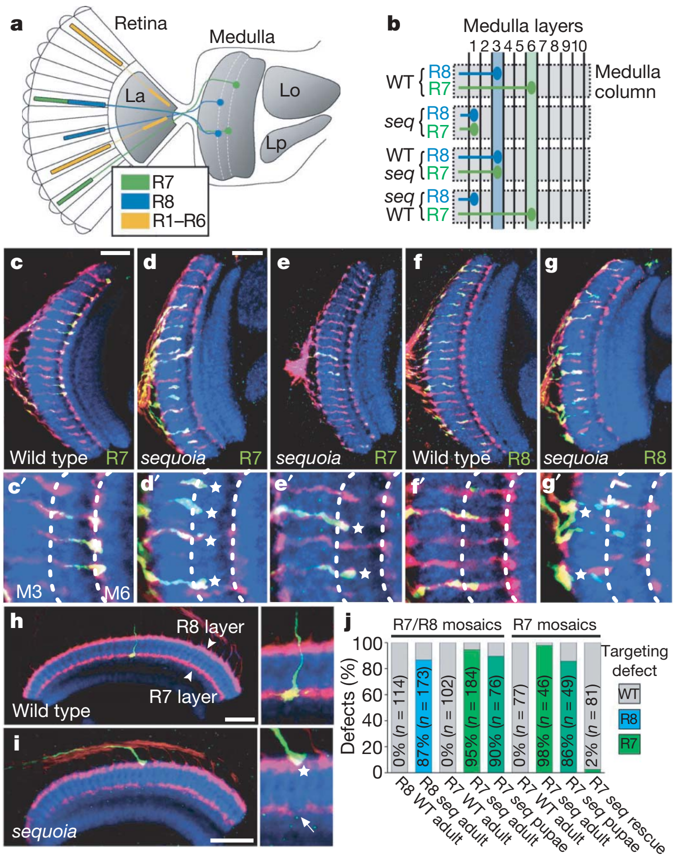


Figure 1 | Mutations in *sequoia* disrupt synaptic-layer selection of R8 and R7 axons. **a**, Schematic representation of *Drosophila* R cell projections from the adult retina into the lamina (La, R1–R6) and medulla (R8/R7) optic ganglia; more central optic ganglia are the lobula (Lo) and the lobula plate (Lp). **b**, The medulla neuropil is organized into vertical layers and horizontal columns. R8 and R7 of an ommatidium project their axons into the same medulla column but synapse in different medulla layers (M3 and M6, respectively). **c**, **f**, Projection of wild-type R7 and R8 axons, in *ey3.5-Flp*-induced, green fluorescent protein (GFP)-labelled MARCM clones, into the M6 and M3 synaptic layer, respectively (**c′**, **f′**, magnified details of **c**, **f**). **d**, **g**, In *seq* mosaics, mutant R7 cells switch from the M6 to the M3 layer (asterisks in **d′**) and mutant R8 cells terminate above the M3 layer (asterisks in **g′**). **e**, Loss of *seq* in a single R7 cell leads to the co-projection of their axon with the neighbouring wild-type R8 cell in M3 (**e′**, magnification of **e**). **h**, **i**, Development of the *seq* mutant axon-targeting defect (**h**, *wild type*; **i**, *sequoia*). Ingrowing *seq* mutant R7 axons terminate in the R8 recipient layer and subsequently become stabilized (40h after puparium formation). Protein labelling in micrographs: anti-GFP (green) for homozygous cells in genetic mosaics, anti-Chaoptin (red) for R8/R7 axons, anti-Cadherin-N (blue) for medulla neuropil. Scale bars, 20 μm. **j**, Quantification of R8/R7 synaptic layer defects in R8/R7-specific and R7-specific *seq* mosaics.

[1]Institut für Neurobiologie, Universität Münster, Badestrasse 9, D-48149 Münster, Germany.

detected in homozygous seq^5 and seq^6 mutant R cells with an anti-sequoia monoclonal antibody (Supplementary Fig. 4) and the R7 targeting defects could be rescued after the expression of a *sequoia* transgene in the mutant cells (Supplementary Fig. 3a). To determine whether *sequoia* functions cell-autonomously in R cells to control axon target selection, we induced R7 cell-type-specific MARCM clones[5]. Almost all (98%, $n = 46$) of the single homozygous mutant R7 terminals surrounded by heterozygous R8 and R7 axons stop prematurely in M3 (Fig. 1b, e, i, j). Because the loss of *sequoia* does not lead to changes in early R8 and R7 specification (Supplementary Fig. 4) or their subsequent projection towards their optic lobe ganglion (Supplementary Fig. 5), we conclude that *sequoia* is required for R8/R7 axons to connect to their synaptic target layer.

The *sequoia* targeting phenotype becomes obvious during early pupal development when the R8 and R7 axons terminate sequentially into two separate medulla layers (Fig. 1h, i, and Supplementary Fig. 6). In contrast to the adult brain, the medulla target field during initial R8/R7 ingrowth appears more homogeneous (Fig. 2d, e) and distinct layer organization becomes visible in the subsequent segregation of lamina and medulla neuron processes[5,10] (M.P., unpublished observations). *seq* mutant R8 and R7 axons reach the medulla target area at the appropriate time and in the correct topographic order, but gaps in the initial R7 layer can be detected; this is due to the termination of R7 axons in the R8 layer (Supplementary Fig. 6a, b, e, f). Similarly to the R7 projection defect in the adult visual system, the mis-targeting of R7 into the R8 layer during the pupal stage seems to be due to the disruption of *seq* function directly in R7 and is not influenced by defects in R8 (Fig. 1h, i, and Supplementary Fig. 6e, f).

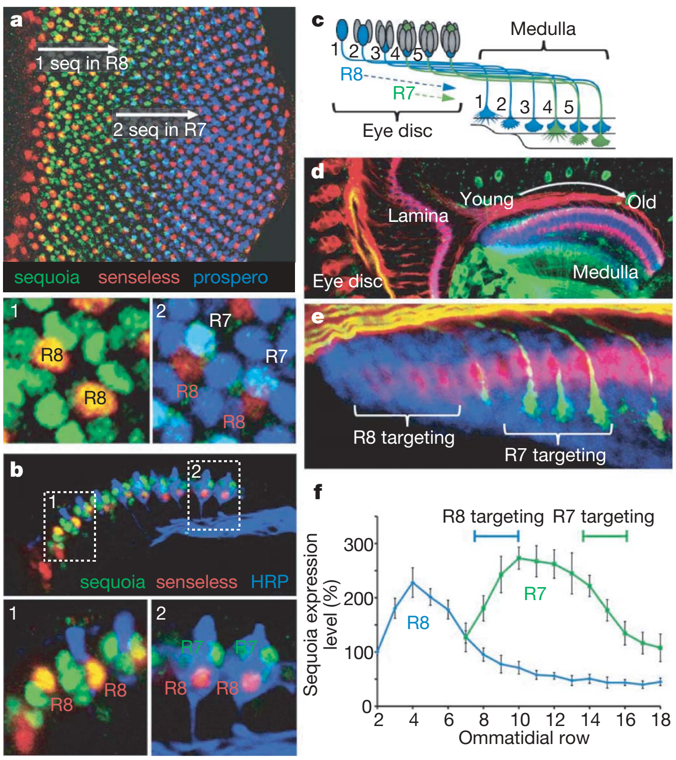


Figure 2 | Transient expression of sequoia in projecting R cells. **a**–**f**, Sequential R8/R7 differentiation in the eye disc (**a**–**c**) and their axonal projections into the medulla (**d**, **e**), shown schematically in **c**. **a**, At the eye disc region of initial neuronal differentiation, sequoia (green) is strongly expressed in R8 cells (senseless-positive, red). In more mature ommatidia, R7 cells start to differentiate and to express sequoia (prospero-positive, blue). **b**, Lateral view of projecting R8/R7 cells showing that the initial sequoia expression in R8 is lost by the time that R7 shows a peak of sequoia expression. **d**, Gradual development of medulla with sequential R8/R7 innervation in the young medulla region, and formation of additional medulla layer in older medulla parts (medulla neurons in green). **e**, Distinct phases of R8 and R7 axon ingrowth and separation (R8 in red, R7 in green). **f**, Correlation of the sequoia expression profiles in R8/R7 and their corresponding phases of axon targeting in the medulla. **a** and **b** show a third-instar eye disc; **d** and **e** show the medulla at 24 h after puparium formation. Error bars in **f** represent s.d.

The expression pattern of sequoia is highly dynamic in early differentiating R8/R7 cells during the phase of axonal growth and medulla innervation (Fig. 2a, b, f, and Supplementary Fig. 7). The onset of sequoia expression in R cells reflects their sequential birth and differentiation order. As a result of the temporal gap in neuronal differentiation between R8 and R7 from the same ommatidium[11], together with a rapid downregulation of sequoia, the periods of elevated sequoia expression in these two R cells are non-overlapping. By the time the R7 cell initiates axonal projection accompanied by a high expression level of sequoia, the axon of the neighbouring R8 cell has already reached the medulla, and sequoia expression has been turned off (Fig. 2a, b). We therefore observed a tight correlation between the R-cell-specific sequoia expression profile and the sequential R8/R7 medulla target layer innervation (Fig. 2e, f).

The fact that sequoia is expressed in both R8 and R7, but the loss of *sequoia* leads to specific R8/R7 axon-targeting defects, raises the possibility that connection specificity is mediated through the difference in the temporal pattern of sequoia expression. To test the importance of precise *sequoia* expression timing, we first synchronized the R8/R7 sequoia expression profiles (Fig. 3). Similarly to the *seq* loss of function situation, constitutive *seq* expression in all R cells does not lead to changes in the cell fate of R8 or R7 (Supplementary Fig. 8a, b) or their initial axonal projection towards the medulla (Supplementary Fig. 9). However, whereas in wild-type individuals R8/R7 growth cones transform their morphology over subsequent hours from an ‘expanded’ to a ‘condensed’ appearance (Fig. 3a–c), growth cones of R8/R7 cells with a prolonged *sequoia* expression remain in the ‘expanded’ state and R8 axons extend towards and terminate precisely in the R7 layer (Fig. 3d–f). The co-innervation of R8 and R7 axons is maintained during the subsequent steps of medulla reorganization[10], because a single R8/R7 layer in M6 can be detected in the medulla of adult flies (Supplementary Fig. 10a, b). These data indicate that the downregulation of *sequoia* expression in R8 cells is critical for their growth cones to become stabilized in their initial target layer. Prolonged *sequoia* expression specifically in R8 or R7 (Supplementary Fig. 10c, d) and the *sequoia* overexpression in R8 in an R7 deficient-background (Fig. 3g, h, and Supplementary Fig. 10e, f) and in single R7 cells (Supplementary Fig. 3) support the model in which the endogenous level of *sequoia* determines the competence of ingrowing R-cell axons to connect to their appropriate target layer (Fig. 3i).

To obtain insights into the molecular mechanisms through which R8/R7 axon targeting competence is mediated, we tested whether the loss of *sequoia* leads to changes in the expression of neuronal cell adhesion/receptor molecules (Fig. 4 and Supplementary Fig. 11). A significant decrease in the expression levels of the receptor protein tyrosine phosphatase Lar (leukocyte-antigen-related-like) (Supplementary Fig. 11a) and the homophilic cell adhesion molecule Cadherin-N (CadN) (Fig. 4a, b) was found in *sequoia* mutant R cells. Although CadN is the only factor known so far to be essential for the initial R8/R7 axon targeting, its widespread expression in R cells and in the target region has made it unlikely that homophilic CadN interactions alone are sufficient to specify R8/R7 target choice[5,10]. We therefore examined whether the *sequoia* targeting mechanism would allow CadN to provide spatio-temporal specificity in initial axon–target adhesion. Loss of *CadN* and *sequoia* leads to an identical early termination of R7 axons in the outer medulla layer and later co-innervation of R7 and R8 in the adult M3 layer (Fig. 4c). After the removal of *CadN* from *sequoia*-overexpressing R8/R7 cells, a complete suppression of the M3→M6 layer switch can be observed because all brains show the characteristic M6→M3 layer switch of *CadN* mutant R7 axons (Fig. 4d, e). This result indicates that the *sequoia*-mediated axon targeting functions through CadN. However, the *sequoia*-induced M3→M6 layer switch is not caused simply by an

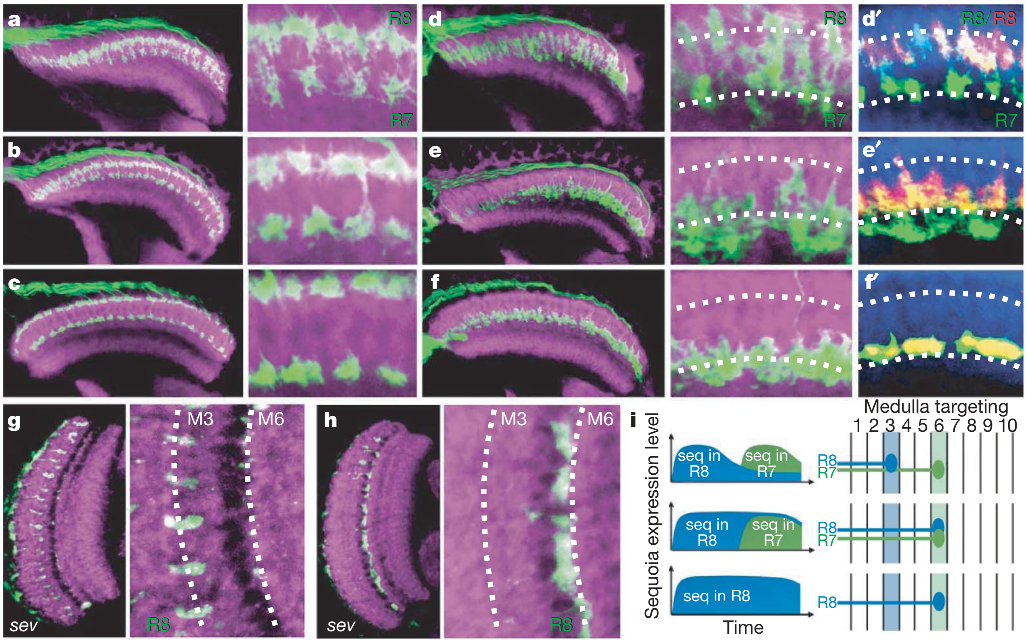


Figure 3 | Temporal control of R8/R7 axon targeting by *sequoia*. **a–c**, Control sequence showing transition of R8/R7 growth cones from an expanded morphology during initial medulla innervation (**a**) through an intermediate appearance (**b**) to a condensed appearance after axon–target stabilization (**c**). **a**, 18 h after puparium formation (APF); **b**, 30 h APF; **c**, 42 h APF. **d–f**, Synchronization of the *sequoia* R8/R7 expression profiles leads to a co-convergence of their axons in the R7 layer; **d**, 18 h APF; **e**, 30 h APF; **f**, 42 h APF. **g**, In R7-cell-deficient (*sevenless*) mutants, R8 axons connect to the M3 layer, whereas the M6 layer remains empty. **h**, Prolonged *sequoia* expression in R8 cells in a *sevenless* background induces an M3→M6 layer switch. Labelling in micrographs: R8/R7 axons are revealed with *GMR–myr–GFP* (GMR, glass multiple reporter) (green) (**a–f**); in **d′–f′** R8 are also labelled with *ato-τ-myc* (red); in **g**, **h** R8 are labelled with anti-Chaoptin (green). Magnifications are × 3. **i**, Schematic illustration of the R8/R7 expression profiles and the corresponding axon-targeting phenotype in wild-type (top row), synchronized *sequoia* expression in R8/R7 (middle row) and prolonged *sequoia* expression in R8 (bottom row).

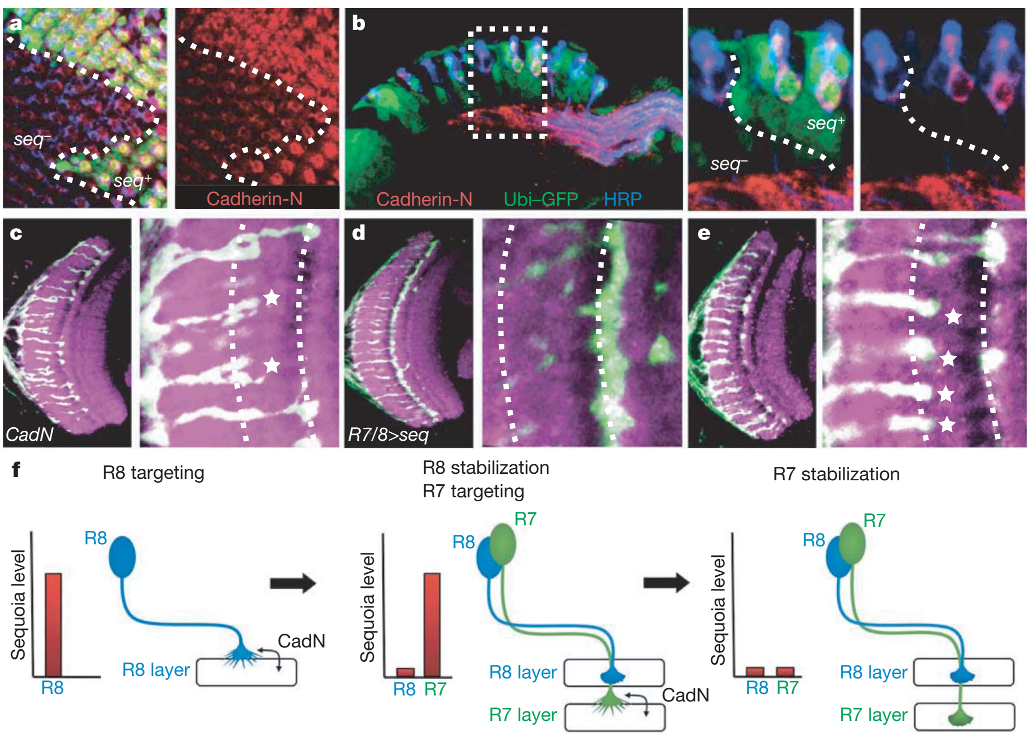


Figure 4 | *sequoia*-induced temporal targeting competence is mediated via Cadherin-N. **a**, **b**, Loss of *sequoia* leads to reduced expression of Cadherin-N in R cells; top view (**a**) and lateral view (**b**). Ubi, *Ubiquitin*; HRP, horseradish peroxidase. **c–e**, *sequoia*-controlled target-layer innervation requires Cadherin-N. **c**, Mutations in *Cadherin-N* (*CadN*) lead to an M6→M3 switch in R7 axon targeting. **d**, *sequoia*-induced M3→M6 switch in R8 targeting. **e**, Loss of Cadherin-N suppresses the *sequoia*-induced M3→M6 switch in R8 targeting. Magnifications are ×3 (**c**) and ×4 (**c–e**). **f**, 'Temporal targeting competence' model of R8/R7 axon layer selection. When the first R8 axons enter the medulla target field they possess a greater targeting competence, allowing them to recognize the primary target layer. This initial recognition is mediated through a direct adhesive interaction through Cadherin-N. By the time that the R7 axon reaches the corresponding medulla column, the target area has developed an additional target layer, which will be recognized by R7, again using Cadherin-N as an adhesive cue. R8 is no longer competent to respond to the newly arising target layer.

increase in CadN adhesion, because overexpression of *CadN* does not affect the R8/R7 projections into the medulla (data not shown) and the overexpression of *sequoia* does not lead to a significant increase in the level of CadN in R cells (Supplementary Fig. 8c). From these data we conclude that homophilic CadN interactions are sequentially used in R8 and R7 to mediate the *sequoia*-regulated axon–target interaction.

On the basis of these results we propose that initial afferent lamination in the developing visual system of *Drosophila* is not controlled by the layer-specific expression of recognition molecules but is mediated through temporally restricted competence of ingrowing axons to interact with the target region (Fig. 4f). Here, permissive cell adhesion molecules provide spatio-temporal recognition specificity in axon–target interaction. Interestingly, Cadherin-N has been shown to function in subsequent steps of medulla innervation by lamina neurons[12], supporting the idea that the same type of adhesion molecule could be used in a repetitive fashion to support sequential layer formation. A critical aspect of the specificity of temporal recognition is the rapid downregulation of the targeting competence after the initial axon–target contact to prevent the reactivation of early projecting axons during the subsequent steps of innervation. The *sequoia*-mediated axon targeting mechanism controls an essential step in the coordinated development between sensory neurons and their central nervous system (CNS) target field in the *Drosophila* visual system. Starting with afferent-derived signals inducing early target neuron differentiation[13,14], the *sequoia*-controlled patterning of initial axon innervation ensures the correct R8/R7 positioning for subsequent afferent-induced target field organization[15] and layer-specific molecule expression to stabilize the initial innervation[16].

Neural progenitors often generate distinct subtypes of neurons in an invariant temporal sequence during development[6,17]. Here we show that, similarly to its role in generating cell type diversity during neuronal division, temporal identity is used subsequently to generate connection diversity in projecting neurons. Mis-expression of *sequoia* in more mature R cells during the later steps of differentiation has no effect on axonal projection (see Supplementary Fig. 10g, h), indicating that R cells lose their competence to respond to the *sequoia*-induced targeting identity. A narrow developmental window of competence has also been described for the temporal identity factors in the embryonic and postembryonic nervous system[17,18]. Similarly, dynamic expression of sequoia can be observed throughout the development of the nervous system in *Drosophila*, and *seq* mutants show a severe disruption of the connectivity pattern in various brain regions (Supplementary Fig. 12, and data not shown). In vertebrates, retina development is also characterized by the sequential generation of distinct cell types followed by their assembly into a highly laminated CNS structure. Recent *in vivo* imaging studies in zebrafish have shown that axon–dendrite interactions occur in a sequential manner and illustrate a transient requirement for some of the cell types during the assembly of laminated retinal circuits[19,20]. Thus, a sequoia-related control mechanism might be more broadly applicable to the development of circuit specificity in the CNS.

METHODS SUMMARY

Mosaic analysis. For mosaic analysis of *sequoia* function in R cells, large clones were generated with the *ey3.5-Flp* construct[21], which is specifically expressed in the eye disc. Homozygous mutant R8 and R7 cells were labelled with the cell-type-specific Gal4 lines *Rh6-Gal4* (ref. 22) (R8), *PANR7-Gal4* (ref. 23) (R7), *LGMR-Gal4* (ref. 24) (R8/R7), *PM181-Gal4* (ref. 5) (R7, pupal stage), or *elav-Gal4* (R8/R7), driving the expression of *UAS-sybGFP*[25], or *UAS-mCD8GFP*[8]. Single homoyzygous R7 cells were generated with *GMR-Flp*[5]. For the *sequoia* rescue experiments we introduced *LGMR-Gal4* into the *GMR-Flp/MARCM* system to drive the expression of *UAS-seq* and *UAS-mCD8GFP*[8].

Expression analysis. The expression dynamics of sequoia in wild-type R8/R7 was analysed in confocal image stacks of stained larval eye discs, and relative expression levels in consecutive ommatidial rows were determined with the ImageJ software. For synchronization of the *sequoia* expression levels in R8/R7 we used the *LGMR-Gal4* line to drive *sequoia* in both cell types, and *MT14-Gal4* (ref. 26) and *PM181-Gal4* (ref. 5) to drive *sequoia* in R8 and R7, respectively. We used *tubGal80*^{ts} (ref. 27) in the genetic scheme to restrict the Gal4-driven *sequoia* overexpression temporally in R8/R7 during development.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 July; accepted 9 September 2008.
Published online 2 November 2008.

1. Sanes, J. R. & Yamagata, M. Formation of lamina-specific synaptic connections. *Curr. Opin. Neurobiol.* **9**, 79–87 (1999).
2. Inoue, A. & Sanes, J. R. Lamina-specific connectivity in the brain: regulation by N-cadherin, neurotrophins, and glycoconjugates. *Science* **276**, 1428–1431 (1997).
3. Yamagata, M., Weiner, J. A. & Sanes, J. R. Sidekicks: synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell* **110**, 649–660 (2002).
4. Gil, O. D., Needleman, L. & Huntley, G. W. Developmental patterns of cadherin expression and localization in relation to compartmentalized thalamocortical terminations in rat barrel cortex. *J. Comp. Neurol.* **453**, 372–388 (2002).
5. Lee, C. H., Herman, T., Clandinin, T. R., Lee, R. & Zipursky, S. L. N-cadherin regulates target specificity in the *Drosophila* visual system. *Neuron* **30**, 437–450 (2001).
6. Isshiki, T., Pearson, B., Holbrook, S. & Doe, C. Q. *Drosophila* neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell* **106**, 511–521 (2001).
7. Ting, C. Y. & Lee, C. H. Visual circuit development in *Drosophila. Curr. Opin. Neurobiol.* **17**, 65–72 (2007).
8. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22**, 451–461 (1999).
9. Brenman, J. E., Gao, F. B., Jan, L. Y. & Jan, Y. N. Sequoia, a tramtrack-related zinc finger protein, functions as a pan-neural regulator for dendrite and axon morphogenesis in *Drosophila. Dev. Cell* **1**, 667–677 (2001).
10. Ting, C. Y. *et al. Drosophila* N-cadherin functions in the first stage of the two-stage layer-selection process of R7 photoreceptor afferents. *Development* **132**, 953–963 (2005).
11. Tomlinson, A. & Ready, D. F. Neuronal differentiation in *Drosophila* ommatidium. *Dev. Biol.* **120**, 366–376 (1987).
12. Nern, A., Zhu, Y. & Zipursky, S. L. Local N-cadherin interactions mediate distinct steps in the targeting of lamina neurons. *Neuron* **58**, 34–41 (2008).
13. Huang, Z. & Kunes, S. Hedgehog, transmitted along retinal axons, triggers neurogenesis in the developing visual centers of the *Drosophila* brain. *Cell* **86**, 411–422 (1996).
14. Huang, Z., Shilo, B. Z. & Kunes, S. A retinal axon fascicle uses spitz, an EGF receptor ligand, to construct a synaptic cartridge in the brain of *Drosophila. Cell* **95**, 693–703 (1998).
15. Bazigou, E. *et al.* Anterograde Jelly belly and Alk receptor tyrosine kinase signaling mediates retinal axon targeting in *Drosophila. Cell* **128**, 961–975 (2007).
16. Shinza-Kameda, M., Takasu, E., Sakurai, K., Hayashi, S. & Nose, A. Regulation of layer-specific targeting by reciprocal expression of a cell adhesion molecule, capricious. *Neuron* **49**, 205–213 (2006).
17. Zhu, S. *et al.* Gradients of the *Drosophila* Chinmo BTB-zinc finger protein govern neuronal temporal identity. *Cell* **127**, 409–422 (2006).
18. Grosskortenhaus, R., Pearson, B. J., Marusich, A. & Doe, C. Q. Regulation of temporal identity transitions in *Drosophila* neuroblasts. *Dev. Cell* **8**, 193–202 (2005).
19. Kay, J. N. *et al.* Transient requirement for ganglion cells during assembly of retinal synaptic layers. *Development* **131**, 1331–1342 (2004).
20. Mumm, J. S. *et al.* In vivo imaging reveals dendritic targeting of laminated afferents by zebrafish retinal ganglion cells. *Neuron* **52**, 609–621 (2006).
21. Chotard, C., Leung, W. & Salecker, I. *glial cells missing* and *gcm2* cell autonomously regulate both glial and neuronal development in the visual system of *Drosophila. Neuron* **48**, 237–251 (2005).
22. Tahayato, A. *et al.* Otd/Crx, a dual regulator for the specification of ommatidia subtypes in the *Drosophila* retina. *Dev. Cell* **5**, 391–402 (2003).
23. Wernet, M. F. *et al.* Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* **440**, 174–180 (2006).
24. Wernet, M. F. *et al.* Homothorax switches function of *Drosophila* photoreceptors from color to polarized light sensors. *Cell* **115**, 267–279 (2003).
25. Blagburn, J. M., Alexopoulos, H., Davies, J. A. & Bacon, J. P. Null mutation in *shaking-B* eliminates electrical, but not chemical, synapses in the *Drosophila* giant fiber system: a structural study. *J. Comp. Neurol.* **404**, 449–458 (1999).
26. Kaminker, J. S., Canon, J., Salecker, I. & Banerjee, U. Control of photoreceptor axon target choice by transcriptional repression of Runt. *Nature Neurosci.* **5**, 746–750 (2002).
27. McGuire, S. E., Mao, Z. & Davis, R. L. Spatiotemporal gene expression targeting with the TARGET and gene-switch systems in *Drosophila. Sci. STKE* **2004**, pl6 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Bellen, J. Brenman, I. Salecker and R. Schuh for fly stocks and antibodies; D. Hawellek for the identification and analysis of *seq*[6]; and H. Aberle, S. Bogdan, C. Klämbt, A. Püschel and members of the Hummel laboratory for critical comments on the manuscript. T.H. is supported by the Deutsche Forschungsgemeinschaft (SFB629) and the EMBO Young Investigator Programme.

Author Contributions M.P. designed and performed all the experiments, and M.P. and T.H. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.H. (hummel@uni-muenster.de).

METHODS

Genetics. Fly stocks were maintained in standard medium at 25 °C unless stated otherwise. The *sequoia* alleles *seq*^{5} and *seq*^{6} were identified in a histological mosaic screen with the *eyFlp/FRT* technique[5] (details of the mutagenesis screen are available from the authors). To label photoreceptor axons and synapses in the adult optic lobe, the following markers were used: *GMR-myr-GFP*[28], *LGMR-Gal4* (ref. 24) to label all R cells, *PANR7-Gal4* (ref. 23) and *Rh6-Gal4* (ref. 22) in combination with *UAS-sybGFP*[25] in MARCM clones for cell-type-specific labelling of R7 and R8. The stocks used were as follows: (1) *ey3.5-Flp; FRT42 seq/CyO*; (2) *ey3.5-Flp; GMR-myr-GFP FRT42 seq/CyO*; (3) *ey3.5-Flp; GMR-myr-GFP FRT42/CyO*; (4) *FRT42 Gal80; PANR7-Gal4 UAS-sybGFP*; (5) *FRT42 Gal80; Rh6-Gal4 UAS-sybGFP* (6) GMR-Flp; FRT42; (7) *GMR-Flp; FRT42 seq*; (8) *FRT42 Gal80; LGMR-Gal4 UAS-mCD8GFP*. For analysis of *sequoia* expression and function during larval and pupal development following stocks were used: (1) *FRT42; ro-τ-lacZ*[29]; (2) *FRT42; ato-τ-myc* (ref. 30); (3) *GMR-Flp; FRT42 seq UAS-mCD8GFP/CyO*; (4) *FRT42 Gal80; elav-Gal4 UAS-mCD8GFP*; (5) *FRT42 Gal80; PM181-Gal4*. The stocks used for the *sequoia* overexpression studies were as follows: (1) *Rh6-EGFP; LGMR-Gal4*; (2) *UAS-seq; Gal80*^{ts}; (3) *sev*^{E1}*; LGMR-Gal4* (ref. 31); (4) *FRT42 Gal80; MT14-Gal4 UAS-mCD8GFP*[26]; (6) *FRT42 Gal80; PM181-Gal4 UAS-lacZ*[5]. Stocks used for the *Cadherin-N* interaction experiments were (1) *ey3.5-Flp; FRT42 seq UAS-CadN/CyO*; (2) *ey3.5-Flp; CadN*^{405} *FRT40* (ref. 5); (3) *Gal80 FRT40; elav-Gal4 UAS-mCDGFP*; (4) *ey3.5-Flp; CadN*^{405} *FRT40 UAS-seq/CyO.*

Immunohistology. Primary antibodies used for immunohistochemistry were: rabbit anti-sequoia[9] (1:1,000 dilution); rat anti-Cadherin-N extracellular domain[32] (DN-Ex no. 8; 1:20; DSHB); rabbit anti-GFP (1:1,000; Molecular Probes); mouse anti-Chaoptin[33] (24B10, 1:40; DSHB); rat anti-Elav[34] (1:10; DSHB); mouse anti-Prospero[35] (1:10; DSHB); guinea pig anti-Senseless[36] (1:1,000); rat anti-Spalt[37] (1:300); mouse anti-DLar[38] (1:10; DSHB); mouse anti-DPTP69D[39] (1:10; DSHB); mouse anti-Neuroglian[40] (1:10; DSHB); mouse anti-Fasciclin II (1:10; DSHB); mouse BP102 (1:10; DSHB); and mouse anti-Flamingo[41] (1:10; DSHB). Secondary antibodies used were as follows (all 1:300, all from Invitrogen): goat anti-rabbit Alexa 488, goat anti-mouse Alexa 488, goat anti-mouse Alexa 568, goat anti-rat Alexa 568, goat anti-guinea pig Alexa 568, goat anti-rabbit Alexa 647 and goat anti-rat Alexa 647. Immunostaining of brains of adult flies and pupae were performed essentially as described in ref. 42 with the following exceptions: (1) adult brains were fixed in 2% paraformaldehyde (PFA) for 90 min, (2) for the dissection of the pupal brains, the pupal cases were opened, 2% PFA was added, and the brains were allowed to fix for 10 min before further dissection in 2% PFA, (3) to avoid cross-reactivity of secondary antibodies in mouse anti-Chaoptin and rat anti-Cadherin-N co-stainings, samples were first stained with mouse anti-Chaoptin and appropriate secondary antibodies, followed by a postfixation in 4% PFA for 30 min and subsequent staining with anti-Cadherin-N in accordance with the standard protocol.

To compare the sequoia expression levels in R8 and R7 during larval development, the mean signal intensity in a defined area was measured for both cells in consecutive ommatidial rows. To correlate the sequoia expression levels with the corresponding targeting events we defined the time required by an R cell to reach its target layer by comparing the amount of differentiated R7/R8 cells in the eye disc with the number of terminals present in the corresponding layer at a given time point, considering that the ommatidial rows develop at a rate of 1.5–2.0 h (ref. 11). Fluorescent samples were analysed with a Zeiss Meta510 confocal microscope and images were processed with Adobe Photoshop.

28. Chang, H. C. *et al.* Hsc70 is required for endocytosis and clathrin function in *Drosophila. J. Cell Biol.* **159,** 477–487 (2002).
29. Garrity, P. A. *et al.* Retinal axon target selection in *Drosophila* is regulated by a receptor protein tyrosine phosphatase. *Neuron* **22,** 707–717 (1999).
30. Senti, K. A. *et al.* Flamingo regulates R8 axon–axon and axon–target interactions in the *Drosophila* visual system. *Curr. Biol.* **13,** 828–832 (2003).
31. Banerjee, U., Renfranz, P. J., Pollock, J. A. & Benzer, S. Molecular characterization and expression of *sevenless*, a gene involved in neuronal pattern formation in the *Drosophila* eye. *Cell* **49,** 281–291 (1987).
32. Iwai, Y. *et al.* Axon patterning requires *D*N-cadherin, a novel neuronal adhesion receptor, in the *Drosophila* embryonic CNS. *Neuron* **19,** 77–89 (1997).
33. Fujita, S. C., Zipursky, S. L., Benzer, S., Ferrus, A. & Shotwell, S. L. Monoclonal antibodies against the *Drosophila* nervous system. *Proc. Natl Acad. Sci. USA* **79,** 7929–7933 (1982).
34. O'Neill, E. M., Rebay, I., Tjian, R. & Rubin, G. M. The activities of two Ets-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* **78,** 137–147 (1994).
35. Campbell, G. *et al.* RK2, a glial-specific homeodomain protein required for embryonic nerve cord condensation and viability in *Drosophila. Development* **120,** 2957–2966 (1994).
36. Nolo, R., Abbott, L. A. & Bellen, H. J. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila. Cell* **102,** 349–362 (2000).
37. Kuhnlein, R. P. *et al. spalt* encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the *Drosophila* embryo. *EMBO J.* **13,** 168–179 (1994).
38. Sun, Q., Bahri, S., Schmid, A., Chia, W. & Zinn, K. Receptor tyrosine phosphatases regulate axon guidance across the midline of the *Drosophila* embryo. *Development* **127,** 801–812 (2000).
39. Desai, C. J., Popova, E. & Zinn, K. A *Drosophila* receptor tyrosine phosphatase expressed in the embryonic CNS and larval optic lobes is a member of the set of proteins bearing the 'HRP' carbohydrate epitope. *J. Neurosci.* **14,** 7272–7283 (1994).
40. Hortsch, M., Bieber, A. J., Patel, N. H. & Goodman, C. S. Differential splicing generates a nervous system-specific form of *Drosophila* neuroglian. *Neuron* **4,** 697–709 (1990).
41. Usui, T. *et al.* Flamingo, a seven-pass transmembrane cadherin, regulates planar cell polarity under the control of Frizzled. *Cell* **98,** 585–595 (1999).
42. Van Vactor, D. L. Jr, Cagan, R. L., Kramer, H. & Zipursky, S. L. Induction in the developing compound eye of *Drosophila*: multiple mechanisms restrict R7 induction to a single retinal precursor cell. *Cell* **67,** 1145–1155 (1991).

LETTERS

Generation of a prostate from a single adult stem cell

Kevin G. Leong[1], Bu-Er Wang[1], Leisa Johnson[1] & Wei-Qiang Gao[1]

The existence of prostate stem cells (PSCs) was first postulated from the observation that normal prostate regeneration can occur after repeated cycles of androgen deprivation and replacement in rodents[1]. Given the critical role of PSCs in maintaining prostate tissue integrity and their potential involvement in prostate tumorigenesis[2], it is important to define specific markers for normal PSCs. Several cell-surface markers have been reported to identify candidate PSCs, including stem cell antigen-1 (Sca-1, also known as Ly6a), CD133 (Prom1) and CD44 (refs 3–10). However, many non-PSCs in the mouse prostate also express these markers and thus identification of a more defined PSC population remains elusive. Here we identify CD117 (c-kit, stem cell factor receptor) as a new marker of a rare adult mouse PSC population, and demonstrate that a single stem cell defined by the phenotype Lin$^-$ Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ can generate a prostate after transplantation *in vivo*. CD117 expression is predominantly localized to the region of the mouse prostate proximal to the urethra and is upregulated after castration-induced prostate involution—two characteristics consistent with that of a PSC marker. CD117$^+$ PSCs can generate functional, secretion-producing prostates when transplanted *in vivo*. Moreover, CD117$^+$ PSCs have long-term self-renewal capacity, as evidenced by serial isolation and transplantation *in vivo*. Our data establish that single cells in the adult mouse prostate with multipotent, self-renewal capacity are defined by a Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ phenotype.

The mouse prostate consists of four pairs of lobes (dorsal, lateral, ventral and anterior lobes) with each divided into three regions relative to the urethra (distal, intermediate and proximal regions; Fig. 1a)[11]. A recent model suggests that the proximal region of prostatic ducts may constitute the stem cell niche in the mouse prostate[12–14]. Using this model, we postulated that new PSC markers would have greater expression in the proximal region relative to the remaining regions of the prostate. We therefore divided wild-type prostates into distal, intermediate and proximal regions and performed quantitative reverse transcriptase polymerase chain reaction (qRT–PCR) analysis for stem cell markers. Four cell-surface

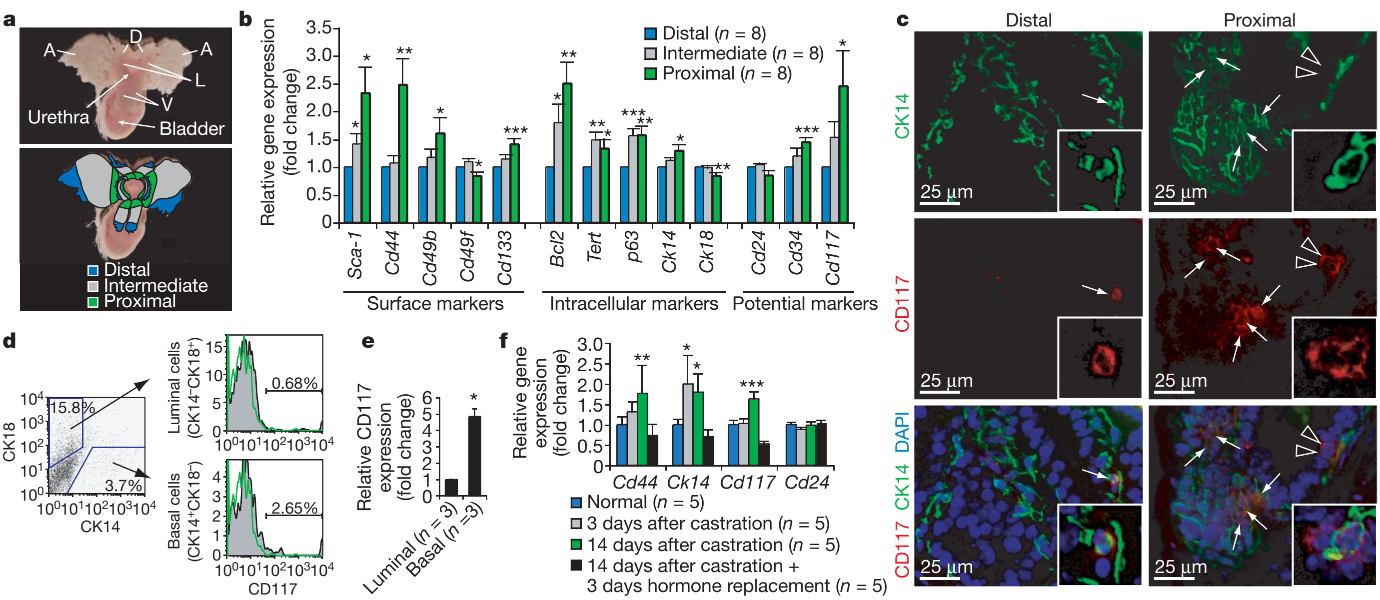


Figure 1 | CD117 is preferentially expressed in the proximal region of the mouse prostate. a, Intact prostate from an adult C57BL/6 mouse. Four pairs of lobes (dorsal (D), lateral (L), ventral (V) and anterior (A)) are shown (top panel). The bottom panel indicates distal, intermediate and proximal regions for each prostatic lobe, relative to the urethra. **b,** qRT–PCR for gene expression in the different regions of the adult C57BL/6 prostate, normalized to the distal region. Statistical comparisons with distal region: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. **c,** Immunofluorescent staining for CK14 (green) and CD117 (red) in distal and proximal regions of an adult C57BL/6 prostate, counterstained with 4,6-diamidino-2-phenylindole (DAPI; blue). Arrows indicate CD117$^+$CK14$^+$ cells and arrowheads indicate CD117$^+$CK14$^-$ cells. Insets show magnified cells. **d,** Flow cytometric analysis of adult C57BL/6 prostate cells triple-stained with CK14, CK18 and CD117, and gated on basal (CK14$^+$CK18$^-$) and luminal (CK14$^-$CK18$^+$) populations. Green line shows control labelling. Dot plots indicate the percentage of basal and luminal cells and histograms indicate the percentage of CD117$^+$ cells in each of the basal and luminal populations. **e,** Quantification of relative CD117 expression in the basal and luminal compartments determined by flow cytometry; $*P = 0.01$. **f,** Microarray analysis of gene expression in adult C57BL/6 prostates (normal, 3 days after castration, 14 days after castration, and 14 days after castration with 3 days of hormone replacement). Data are expressed as the fold change relative to normal. Statistical comparisons with hormone replacement: $*P < 0.05$; $**P < 0.01$; $***P < 0.0001$. All bars represent the mean and s.e.m.

[1]Department of Molecular Biology, Genentech, Inc., 1 DNA Way, South San Francisco, California 94080, USA.

markers (Sca-1, CD44, CD49b (also known as Itga2) and CD133) all of which are known markers of PSCs[3–8], and three intracellular stem-cell markers (Bcl2, telomerase reverse transcriptase (Tert) and p63) showed preferential expression in the proximal region (Fig. 1b and Supplementary Fig. 1a), thus confirming the validity of this assay. The fact that CD44, CD49b, CD133, Bcl2, Tert and p63 are all prostatic basal markers[4–6,8,15,16] suggests that basal markers, relative to luminal markers, may be expressed at increased levels in the proximal region. Consistently, the basal marker cytokeratin 14 (CK14, also known as Krt14) but not the luminal marker CK18 (Krt18) showed preferential expression in the proximal region (Fig. 1b). These data support the previous suggestion that PSCs may constitute a subpopulation of basal cells[8,9].

We next assessed the expression levels of stem cell-surface markers that were not previously reported to identify normal PSCs. CD34 and CD117, but not CD24, were predominantly expressed in the proximal region (Fig. 1b). Because CD117 showed greater differential expression between the proximal and distal regions compared to CD34 (Fig. 1b), we focused on CD117 as a potential PSC marker. Immunostaining confirmed a basal $CD117^+CK14^+$ population with a predominantly proximal expression pattern (Fig. 1c). A proximal $CD117^+CK14^-$ population, however, was also observed (Fig. 1c) with subsequent analysis identifying a luminal $CD117^+CK18^+$ population (Supplementary Fig. 1b). We next performed flow cytometry with triple labelling for CK14, CK18 and CD117 (Fig. 1d). Although $CD117^+$ cells were enriched in the basal compartment (Fig. 1e) these findings indicated that $CD117^+$ cells were not exclusively localized to either the basal or luminal compartments. Moreover, CD117 expression was not confined to a particular prostatic lobe. Instead, CD117—along with CD44, CD49b and CD133—was expressed in all four pairs of lobes, with prominent expression detected in the dorsal prostate (Supplementary Fig. 1c, d). CD117 therefore has an expression profile that is similar to that of known PSC markers.

Although normal PSCs are androgen-independent and survive castration, they remain androgen-responsive and effect prostate regeneration after hormone replacement[1]. If normal PSCs expressed CD117, we postulated that CD117 expression would increase after castration (owing to stem-cell enrichment) and decrease after hormone replacement (owing to differentiated cell expansion). Indeed, CD117, CK14 and CD44, but not CD24, showed this pattern (Fig. 1f and Supplementary Fig. 1e), further indicating that the expression of CD117 is compatible with that of a normal PSC marker.

To provide functional evidence that the $CD117^+$ population is enriched for PSCs, we prepared $CD117^{+/-}$ fractions (Supplementary Fig. 2) and performed prostate colony formation assays *in vitro*[13]. $CD117^+$ cells (but not $CD117^-$ cells) gave rise to several lumen-containing colonies (Fig. 2a). Although this *in vitro* assay suggests that the $CD117^+$ population contains PSCs, the ability of $CD117^+$ cells to generate prostates *in vivo* is an essential assessment of the stem-cell phenotype. Using an *in vivo* prostate generation system[17,18] $CD117^{+/-}$ fractions from C57BL/6 mouse donors were combined with rat embryonic urogenital sinus mesenchymal (UGM) stromal cells and implanted under the renal capsule of athymic nu/nu mouse hosts. Although $CD117^-$ cells remained viable under the renal capsule (Supplementary Fig. 3), grafts of $CD117^-$ cells were small, opaque and similar to grafts of UGM cells alone in their inability to generate prostates (prostate generation frequency (pgf) = 1/10; Fig. 2b, c). In contrast, $CD117^+$ grafts were large, vascularized and translucent (pgf = 10/12; Fig. 2b, c). Histological examination of $CD117^+$ grafts demonstrated a branching morphology (Fig. 2d) with epithelial tubules composed of basal (CK14; Fig. 2e) and luminal (CK18; Fig. 2e) cell lineages. Rare neuroendocrine cells—identified as solitary synaptophysin $(Syp)^+$ cells in the basal compartment of wild-type mouse prostates (Supplementary Fig. 4)[19]—were observed in several prostatic ducts and acini, within the same and across multiple implants (Fig. 2f). $CD117^+$ grafts also expressed the prostate-specific proteins probasin (Pbsn)[20] (Fig. 2g) and Nkx3-1 (ref. 21; Fig. 2h), indicating functional prostate generation. Using a mouse β1 integrin (also known as Itgb1)-specific antibody, we verified that $CD117^+$ grafts were of mouse origin and not due to contaminating rat epithelial cells from the UGM stromal cell preparations (Fig. 2i). Furthermore, we confirmed that the generated prostates were derived from transplanted $CD117^+$ donor cells using an MHC class I

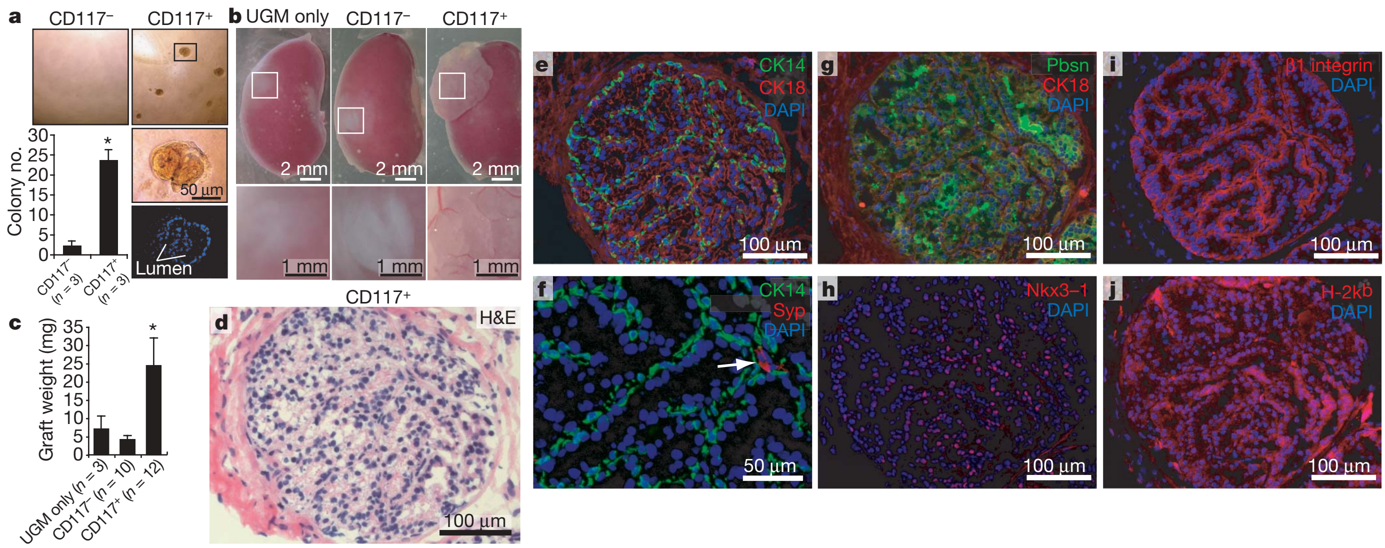


Figure 2 | The $CD117^+$ population enriches for normal adult prostate stem cells. **a**, Colony forming abilities of $CD117^-$ and $CD117^+$ cell populations from adult C57BL/6 prostates. The boxed area is shown enlarged in the middle right image, and a DAPI-stained confocal image demonstrating lumen formation is shown in the bottom right image. Data are from three independent experiments; $*P < 0.0001$. **b**, Prostate generation capacities of UGM only, $CD117^-$ and $CD117^+$ implants 2 months after renal capsule implantation. The boxed areas in the top panels are shown enlarged in the bottom panels. **c**, Quantification of graft weight 2 months after renal capsule implantation. Data are from five independent experiments. *, $CD117^+$ versus $CD117^-$: $P = 0.02$; $CD117^+$ versus UGM only: $P = 0.03$. Data are mean and s.e.m. **d**–**j**, Histological analyses of serial sections of a $CD117^+$ graft stained for (**d**) haematoxylin and eosin (H&E), (**e**) CK14 (green), CK18 (red) and DAPI (blue), (**f**) CK14 (green), Syp (red) and DAPI (blue), (**g**) Pbsn (green), CK18 (red) and DAPI (blue), (**h**) Nkx3-1 (red) and DAPI (blue), (**i**) β1 integrin (red) and DAPI (blue), (**j**) $H\text{-}2k^b$ (red) and DAPI (blue). Arrow indicates a Syp^+ cell.

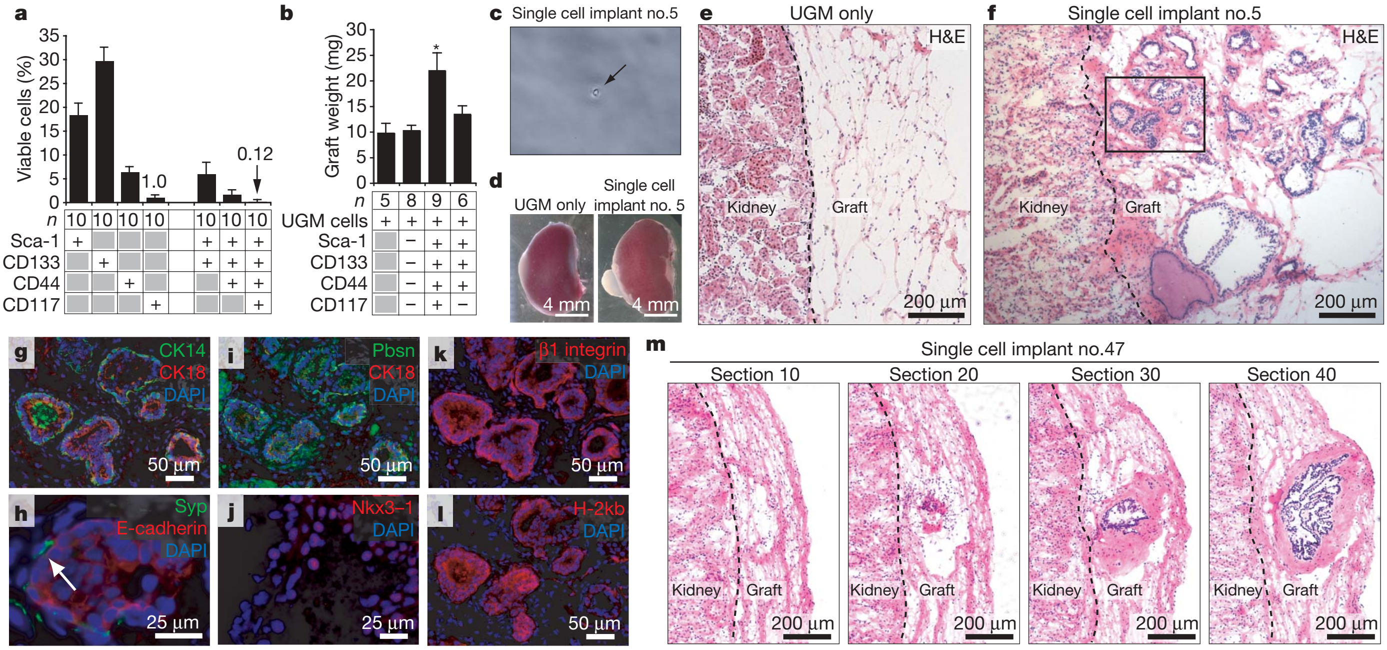


Figure 3 | A single adult stem cell with the phenotype Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ can generate a secretion-producing prostate. **a**, Percentage of viable cells in the adult C57BL/6 prostate expressing single and multiple markers of prostate stem cells. **b**, Quantification of graft weight three months after renal capsule implantation. Data are from two independent experiments. *, Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ versus Lin$^-$Sca-1$^-$CD133$^-$CD44$^-$CD117$^-$: $P = 0.002$; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ versus Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^-$: $P = 0.03$; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ versus UGM only: $P = 0.004$. **c**, Image of the single viable Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ cell used for single cell implant no. 5 (indicated by an arrow). **d**, Prostate generation capacities of UGM only and Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ single cell implant no. 5 three months after renal capsule implantation. **e**, Haematoxylin and eosin (H&E) stain of a UGM-only graft. Dotted line indicates the edge of the kidney parenchyma. **f**, Haematoxylin and eosin stain of single cell implant no. 5. Dotted line indicates the edge of the kidney parenchyma. **g–l**, Serial sections of the boxed area are shown stained for (**g**) CK14 (green), CK18 (red) and DAPI (blue), (**h**) Syp (green), E-cadherin (also known as Cdh1; red) and DAPI (blue; arrow indicates a Syp$^+$ cell), (**i**) Pbsn (green), CK18 (red) and DAPI (blue), (**j**) Nkx3-1 (red) and DAPI (blue), (**k**) β1 integrin (red) and DAPI (blue), and (**l**) H-2k^b (red) and DAPI (blue). **m**, Haematoxylin and eosin stain of single cell implant no. 47. Serial sections 10, 20, 30 and 40 are shown. Dotted line indicates the edge of the kidney parenchyma. All bars represent the mean and s.e.m.

haplotype H-2k^b antibody (Fig. 2j) that specifically recognizes donor (C57BL/6) but not host (athymic nu/nu) mouse cells (Supplementary Fig. 5). These findings demonstrate that the CD117$^+$ population is enriched for normal PSCs with functional prostate generation capacity.

To evaluate the self-renewal capacity of CD117$^+$ cells—a defining characteristic of stem cells[22]—and to determine whether decreased amounts of CD117$^+$ cells would retain prostate generation capacity, we performed serial transplantations with successively reduced numbers of CD117$^+$ cells (Supplementary Fig. 6a). Secondary and tertiary transplants of CD117$^+$ cells, but not CD117$^-$ cells, gave rise to functional prostates comprised of multiple cell types derived from donor C57BL/6 mouse cells (Supplementary Fig. 6b–q). These findings provide direct evidence that the CD117$^+$ population contains normal PSCs with self-renewal capacity.

We next determined whether CD117 signalling was important for normal prostate development. Mouse prostate development begins with epithelial budding from the urogenital sinus at 17.5 days of gestation, with extensive ductal outgrowth and branching occurring during the first 3 weeks of postnatal development[11]. Mice homozygous for the dominant-white spotting locus (W/W) lack CD117 signalling and are perinatal lethal[23], thus precluding an assessment of prostate development. We therefore analysed heterozygous W/W^v mice, which are viable and have partially impaired CD117 signalling[23]. At 4 weeks of age, despite an equivalent body size the mutant prostates were decreased in size with a similar reduction in adulthood (Supplementary Fig. 7a–d). We next examined cell proliferation, differentiation and survival in mutant prostates, because CD117 signalling regulates these processes in various stem cells[24]. Mutant prostates showed inhibited proliferation, although basal and luminal differentiation, cell survival and vascular/smooth muscle cell recruitment were unaltered (Supplementary Fig. 7e–h). Similarly, wild-type C57BL/6 prostates cultured *ex vivo* in the presence of a function-blocking anti-CD117 antibody (ACK2) showed inhibited growth and reduced branching (Supplementary Fig. 8a–c). We confirmed that ACK2 inhibited CD117 signalling by assessing the expression of Slug (also known as Snai2; Supplementary Fig. 8d–f)—a downstream target of the CD117 pathway[25]. Notably, treated prostates showed attenuated proliferation and an increased basal-to-luminal cell ratio, with no effect on cell survival (Supplementary Fig. 8g–l). To evaluate further a possible role for CD117 in PSC function *in vivo*, we administered ACK2 to castrated C57BL/6 mice at an *in vivo* inhibitory dose[26] and assessed prostate regeneration after hormone replacement. Attenuating CD117 function *in vivo* inhibited prostate regeneration concordant with inhibited proliferation and an increased basal-to-luminal cell ratio, with no effect on cell survival or vascular/smooth muscle cell recruitment (Supplementary Fig. 9a–i). These findings suggest that the impairment of CD117 signalling using antagonistic blockers, in contrast to the partial impairment as seen in W/W^v mice, may inhibit prostatic luminal cell differentiation and highlight a potential role for CD117 signalling in normal prostate development. Because CD117 signalling is important for bone-marrow-derived cellular function (including haematopoietic stem and endothelial progenitor cell mobilization[27]) and the vasculature and its supporting stroma may establish a PSC niche[14], it is possible that abrogated CD117 function may adversely affect non-epithelial cell recruitment/maintenance in the prostate, which in turn could compromise prostate development. Although we did not observe deficient vascular/smooth muscle cell recruitment in CD117 mutant and ACK2-treated mice, our studies do not rule out possible effects of abrogated CD117 signalling on non-epithelial cells in the prostate. Studies investigating the prostate-specific

conditional ablation of CD117 are needed to demonstrate conclusively a functional requirement for CD117 in normal prostate development.

To compare the percentage of CD117$^+$ cells in the prostate with that of other PSC populations, we obtained a lineage-depleted (Lin$^-$) population and performed flow cytometry. Whereas CD117$^+$ cells had the lowest frequency in the viable cell population (~1%), Sca-1$^+$ and CD133$^+$ cells were detected at much higher frequencies (Fig. 3a). Indeed, other studies have reported Sca-1 and CD133 expression in both stem and non-stem-cell types, including stromal and differentiated epithelial cells[7,28]. These higher frequencies suggest that Sca-1 and CD133 may mark heterogeneous subpopulations of prostate cells. Given the heterogeneity of single-stained cell populations, we would not expect each marker used alone to yield a subpopulation composed entirely of stem cells. We therefore performed fluorescence-activated cell sorting (FACS) to obtain cell populations expressing combinations of multiple surface markers (Supplementary Fig. 10), followed by renal capsule implantation. We determined that PSCs resided exclusively within the Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ population, which constituted 0.12% of the viable cell population (pgf: Lin$^-$Sca-1$^-$CD133$^-$CD44$^-$CD117$^-$, 0/8; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$, 6/9; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^-$, 0/6; Fig. 3a, b and Supplementary Fig. 11a–h). Serial transplantation yielded similar results (pgf: Lin$^-$Sca-1$^-$CD133$^-$CD44$^-$CD117$^-$, 0/3; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$, 1/3; Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^-$, 0/3; Supplementary Fig. 12a–f). Hence the Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ population contains normal PSCs with self-renewal capacity.

To definitively prove that prostate generation could be achieved from a single Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ cell, we sorted single viable cells into individual wells, imaged each well to confirm the presence of a single cell, and grafted single donor C57BL/6 mouse cells in combination with rat UGM stromal cells under the renal capsule of host athymic nu/nu mice (Supplementary Fig. 13). A total of 14 prostates were generated from 97 single cell transplants (Fig. 3c, d). Histological analyses confirmed that whereas grafts of UGM cells alone were incapable of prostate generation (Fig. 3e), grafts of the 14 successful single cell transplants showed substantial prostate development (Fig. 3f). We confirmed the presence of epithelial tubules comprised of multiple cell lineages (Fig. 3g, h) as well as the expression of Pbsn (Fig. 3i) and Nkx3-1 (Fig. 3j). Notably, single-cell-derived prostates expressed mouse-specific β1 integrin (Fig. 3k) and C57BL/6 donor-specific H-2k^b (Fig. 3l). We further confirmed the donor origin of the generated prostates by PCR-based genotyping of laser capture microdissected cells (Supplementary Fig. 14a). Single-cell-derived prostates had an interconnected branching morphology surrounded by a layer of stromal cells and connective tissue (Fig. 3m). By limiting dilution analysis, the frequency of PSCs in the Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ population was calculated as 1 in 10 (Supplementary Fig. 14b).

To determine whether CD117 would also mark a potential PSC population in the human prostate, we performed flow cytometric analysis of human benign prostatic hyperplasia (BPH; $n = 5$) and benign non-BPH ($n = 4$) specimens. CD117$^+$ cells were observed at a low frequency in the viable cell population in benign non-BPH and BPH specimens (~0.2% and 0.4%, respectively; Fig. 4a, b). By combining multiple stem cell markers, we determined the frequency of viable CD133$^+$CD44$^+$CD117$^+$ cells (a Sca-1 human orthologue has not been identified) in benign non-BPH and BPH specimens to be 0.004% and 0.01%, respectively (Fig. 4a, b). Immunostaining revealed CD117$^+$ cells in the prostate epithelium that co-expressed the basal marker p63 in both benign non-BPH (Fig. 4c–e) and BPH (Fig. 4f–h) specimens. These findings suggest that CD117 expression, in addition to marking a PSC population in the mouse prostate, may also mark a potential PSC population in the human prostate. *In vivo* prostate generation from human CD117$^+$ prostate cells, however, remains to be demonstrated.

This is, to our knowledge, the first report to demonstrate prostate generation from a single adult stem cell. Recently, two studies described mouse mammary gland reconstitution from a single stem cell[29,30]. By demonstrating that a functional organ could be generated from a single adult stem cell, these studies signified hallmark advancement in the stem cell research field. Our study now adds to this advancement. Determining the phenotype of stem cells with single cell tissue generation capacity has important implications for

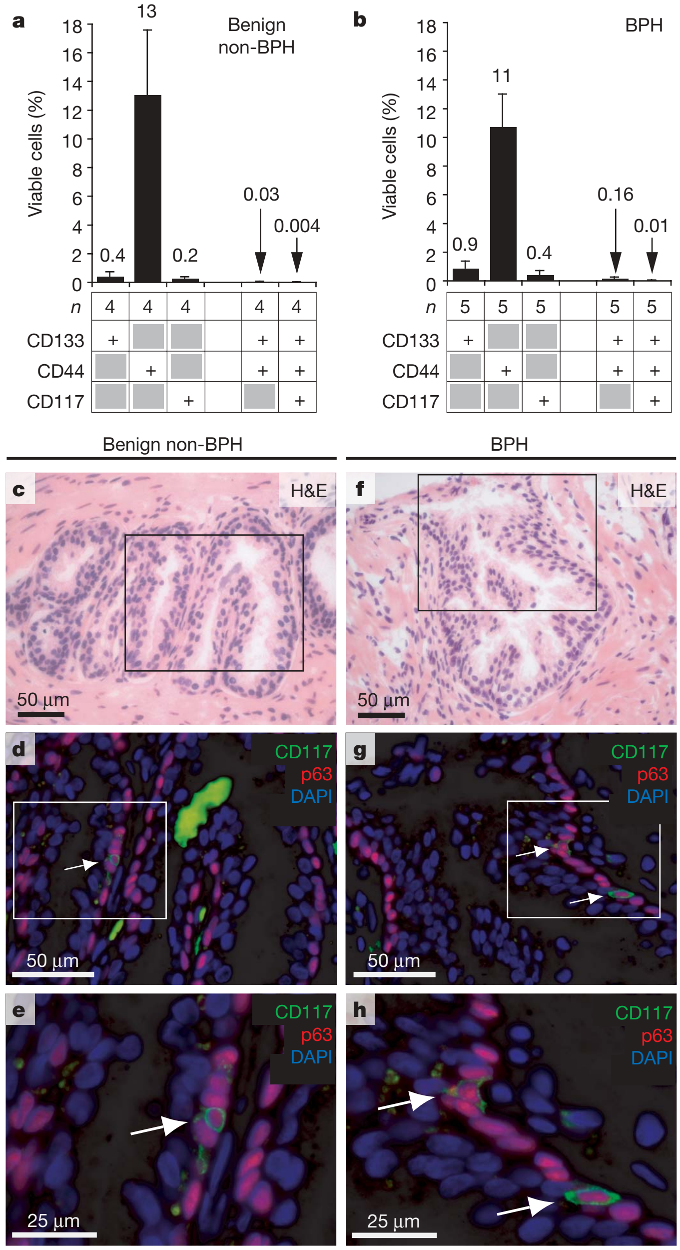


Figure 4 | Human prostates contain a subpopulation of CD117$^+$ cells that localize to the basal layer of the prostate epithelium. **a**, **b**, Percentage of viable cells in human clinical benign non-BPH prostate specimens (**a**) or human clinical BPH prostate specimens (**b**) expressing single and multiple markers of prostate stem cells. **c–h**, Histological analyses of a benign non-BPH prostate specimen (**c–e**) and a BPH prostate specimen (**f–h**) stained for haematoxylin and eosin (H&E) (**c**, **f**) and CD117 (green), p63 (red) and DAPI (blue) (**d**, **e**, **g**, **h**). The boxed areas in **c** and **f** are shown enlarged in **d** and **g**, respectively, and the boxed areas in **d** and **g** are shown enlarged in **e** and **h**, respectively. Arrows indicate cells co-expressing CD117 and p63. All bars represent the mean and s.e.m. The benign non-BPH and BPH specimens used for histological analyses were paired specimens taken from the same patient.

tissue repair and regrowth and for cancer stem/initiating cell identification.

METHODS SUMMARY

Prostate generation *in vivo*. The prostate generation assay was described previously[17,18]. For primary transplants of $CD117^+$ cells, prostates from 8–10-week-old C57BL/6 mice were dissociated and magnetically sorted with anti-mouse CD117 microbeads (Miltenyi Biotec) into $CD117^{+/-}$ fractions (~500,000 dissociated cells were obtained per prostate, with $CD117^+$ cells constituting 7% of magnetically sorted cells, of which $17.5 \pm 2.4\%$ ($n = 10$) were viable as determined by flow cytometry). Sorted cells (100,000 cells per graft) were mixed with UGM stromal cells (250,000 cells per graft) in 3 mg ml^{-1} collagen type I (20 μl per graft), incubated at 37 °C for 1 h to allow collagen gelation, and overlaid with prostate culture medium. After incubation overnight at 37 °C, collagen gels were grafted under the renal capsule of 6–8-week-old athymic nu/nu mice, along with a subcutaneous 90-day slow-release testosterone pellet (12.5 mg per pellet per mouse; Innovative Research of America). Grafts were collected 8 weeks after implantation. For primary transplants of $Lin^-Sca\text{-}1^+CD133^+CD44^+CD117^+$ cells, prostates were sorted by FACS, and the sorted cells (1,300 cells per graft) were mixed with UGM stromal cells (250,000 cells per graft). Grafts were collected 12 weeks after implantation.

Single cell FACS and prostate generation *in vivo*. Details of the procedure are described in Supplementary Fig. 13. Prostates from 8–10-week-old C57BL/6 mice were dissociated and lineage-depleted using a Mouse Lineage Cell Depletion Kit (Miltenyi Biotec), along with a biotin-conjugated anti-mouse CD31 monoclonal antibody (BD Biosciences; clone 390). FACS was performed with a FACSAria flow cytometer (Becton Dickinson). Compensation adjustments were performed with single colour positive controls. Single cells were sorted into Microtest U-bottom 96 well plates (BD Falcon) containing 20 μl collagen type I at 3 mg ml^{-1}. A total of 127 individual wells from single cell FACS were examined, with 106 wells verified to contain a single viable cell from six independent experiments.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 January; accepted 16 September 2008.
Published online 22 October 2008.

1. Isaacs, J. T. in *Control of Cell Proliferation and Cell Death in the Normal and Neoplastic Prostate. Benign Prostatic Hyperplasia* (eds C. H. Rodgers *et al.*) Vol. II 85–94 (National Institutes of Health, Report No. 87–2881, 1987).
2. Abate-Shen, C. & Shen, M. M. Molecular genetics of prostate cancer. *Genes Dev.* **14,** 2410–2434 (2000).
3. Burger, P. E. *et al.* Sca-1 expression identifies stem cells in the proximal region of prostatic ducts with high capacity to reconstitute prostatic tissue. *Proc. Natl Acad. Sci. USA* **102,** 7180–7185 (2005).
4. Liu, A. Y. *et al.* Cell–cell interaction in prostate gene regulation and cytodifferentiation. *Proc. Natl Acad. Sci. USA* **94,** 10705–10710 (1997).
5. Richardson, G. D. *et al.* CD133, a novel marker for human prostatic epithelial stem cells. *J. Cell Sci.* **117,** 3539–3545 (2004).
6. Tsujimura, A. *et al.* Prostatic stem cell marker identified by cDNA microarray in mouse. *J. Urol.* **178,** 686–691 (2007).
7. Xin, L., Lawson, D. A. & Witte, O. N. The Sca-1 cell surface marker enriches for a prostate-regenerating cell subpopulation that can initiate prostate tumorigenesis. *Proc. Natl Acad. Sci. USA* **102,** 6942–6947 (2005).
8. Collins, A. T., Habib, F. K., Maitland, N. J. & Neal, D. E. Identification and isolation of human prostate epithelial stem cells based on $\alpha_2\beta_1$-integrin expression. *J. Cell Sci.* **114,** 3865–3872 (2001).
9. Lawson, D. A., Xin, L., Lukacs, R. U., Cheng, D. & Witte, O. N. Isolation and functional characterization of murine prostate stem cells. *Proc. Natl Acad. Sci. USA* **104,** 181–186 (2007).
10. Schmelz, M. *et al.* Identification of a stem cell candidate in the normal human prostate gland. *Eur. J. Cell Biol.* **84,** 341–354 (2005).
11. Sugimura, Y., Cunha, G. R. & Donjacour, A. A. Morphogenesis of ductal networks in the mouse prostate. *Biol. Reprod.* **34,** 961–971 (1986).
12. Salm, S. N. *et al.* TGF-β maintains dormancy of prostatic stem cells in the proximal region of ducts. *J. Cell Biol.* **170,** 81–90 (2005).
13. Tsujimura, A. *et al.* Proximal location of mouse prostate epithelial stem cells: a model of prostatic homeostasis. *J. Cell Biol.* **157,** 1257–1265 (2002).
14. Wang, G. M., Kovalenko, B., Wilson, E. L. & Moscatelli, D. Vascular density is highest in the proximal region of the mouse prostate. *Prostate* **67,** 968–975 (2007).
15. Bonkhoff, H. & Remberger, K. Differentiation pathways and histogenetic aspects of normal and abnormal prostatic growth: a stem cell model. *Prostate* **28,** 98–106 (1996).
16. Signoretti, S. *et al.* p63 is a prostate basal cell marker and is required for prostate development. *Am. J. Pathol.* **157,** 1769–1775 (2000).
17. Cunha, G. R. & Lung, B. The possible influence of temporal factors in androgenic responsiveness of urogenital tissue recombinants from wild-type and androgen-insensitive (Tfm) mice. *J. Exp. Zool.* **205,** 181–193 (1978).
18. Xin, L., Ide, H., Kim, Y., Dubey, P. & Witte, O. N. *In vivo* regeneration of murine prostate from dissociated cell populations of postnatal epithelia and urogenital sinus mesenchyme. *Proc. Natl Acad. Sci. USA* **100,** 11896–11903 (2003).
19. Kurita, T., Medina, R. T., Mills, A. A. & Cunha, G. R. Role of p63 and basal cells in the prostate. *Development* **131,** 4955–4964 (2004).
20. Johnson, M. A., Hernandez, I., Wei, Y. & Greenberg, N. Isolation and characterization of mouse probasin: An androgen-regulated protein specifically expressed in the differentiated prostate. *Prostate* **43,** 255–262 (2000).
21. Shen, M. M. & Abate-Shen, C. Roles of the *Nkx3.1* homeobox gene in prostate organogenesis and carcinogenesis. *Dev. Dyn.* **228,** 767–778 (2003).
22. Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* **100,** 157–168 (2000).
23. Lev, S., Blechman, J. M., Givol, D. & Yarden, Y. Steel factor and c-*kit* protooncogene: genetic lessons in signal transduction. *Crit. Rev. Oncog.* **5,** 141–168 (1994).
24. Ashman, L. K. The biology of stem cell factor and its receptor C-kit. *Int. J. Biochem. Cell Biol.* **31,** 1037–1051 (1999).
25. Perez-Losada, J. *et al.* Zinc-finger transcription factor Slug contributes to the function of the stem cell factor c-kit signaling pathway. *Blood* **100,** 1274–1286 (2002).
26. Moss, K. G., Toner, G. C., Cherrington, J. M., Mendel, D. B. & Laird, A. D. Hair depigmentation is a biological readout for pharmacological inhibition of KIT in mice and humans. *J. Pharmacol. Exp. Ther.* **307,** 476–480 (2003).
27. Heissig, B., Werb, Z., Rafii, S. & Hattori, K. Role of c-kit/Kit ligand signaling in regulating vasculogenesis. *Thromb. Haemost.* **90,** 570–576 (2003).
28. Shmelkov, S. V. *et al.* CD133 expression is not restricted to stem cells, and both $CD133^+$ and $CD133^-$ metastatic colon cancer cells initiate tumors. *J. Clin. Invest.* **118,** 2111–2120 (2008).
29. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439,** 84–88 (2006).
30. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439,** 993–997 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Cupp, L. Gilmour and W. Tombo for FACS support, X.-D. Wang, R. Soriano and Z. Modrusan for microarray services, I. Kasman, L. Komuves and J. Eastham-Anderson for microscopy assistance, B. D. Tarlow for assistance with laser-capture microdissection and PCR-based genotyping, Genentech Laboratory Animal Resources for support services, L. Blocher and J. Simko for human clinical prostate specimens, and F. J. de Sauvage for valuable discussion, input and supervision of the project.

Author Contributions K.G.L. planned the project, designed and performed experiments, and prepared the manuscript. B.-E.W. assisted with *ex vivo* prostate studies. L.J. provided input and supervised the project. W.-Q.G. conceptualized the study, planned and supervised the project. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to W.-Q.G. (gao@gene.com).

METHODS

Animals. Pregnant SD rats and C57BL/6 male mice (postnatal day 4 and 8–10-weeks-old) were purchased from Charles River Laboratories, athymic nu/nu male mice (6–8-weeks-old) were purchased from Harlan Sprague Dawley, and WBB6F1/J male mice (wild type or W/W^v; 4–8-weeks-old) were purchased from the Jackson Laboratory. The W allele encodes a *Cd117* gene with a deletion of the transmembrane domain and the amino terminus of the kinase domain, whereas the W^v allele encodes a *Cd117* gene with a single point mutation. All experiments were approved by the Animal Research Ethics Committee of Genentech.

Human prostate clinical specimens. Freshly resected human prostate specimens (both BPH and benign non-BPH specimens, distinguished by gross examination by a pathologist; wet weights between 1 and 3 g) were obtained from Bio-options Inc. and The University of California, San Francisco, from consenting patients in accordance with federal and state guidelines.

Antibodies. Antibodies were purchased from the following sources: BD Biosciences, allophycocyanin (APC)-conjugated CD117 (anti-mouse: clone 2B8; anti-human: clone YB5.B8), phycoerythrin (PE)-Cy7-conjugated Sca-1 (clone D7), Ki67 (clone B56), E-cadherin (clone 36), active caspase 3 (polyclonal 557035); eBioscience, PE-conjugated CD133 (anti-mouse: clone 13A4), APC-Alexa Fluor 750-conjugated CD44 (anti-mouse/human: clone IM7), function-blocking CD117 (clone ACK2); Miltenyi Biotec, PE-conjugated CD133 (anti-human: clone AC133); Abcam, CK18 (clone C-04), H-2k^b (clone ER-HR52), CD117 (polyclonal ab956); Chemicon, mouse-specific β1 integrin (clone MB1.2), synaptophysin (clone SY38); R&D Systems, CD117 (clone 180627); Covance, CK14 (polyclonal AF64); AbD Serotec, CD31 (clone 2H8); Sigma, α-smooth muscle actin (SMA) (clone 1A4); Santa Cruz Biotechnology, probasin (polyclonal M-18), p63 (clone 4A4); Invitrogen, synaptophysin (polyclonal Z66), secondary antibodies conjugated to Alexa Fluor 488 or 594. Nkx3-1 polyclonal antibody was a gift from C. Abate-Shen.

UGM stromal cell preparation. The UGM isolation procedure has been described previously[18]. In brief, E18 embryos from pregnant SD rats were killed and urogenital sinuses were collected. After separation of the UGM from the urogenital sinus epithelium, the UGM was digested with 1 mg ml^{-1} collagenase/dispase (Roche) in DMEM supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 U ml^{-1} penicillin and 100 mg ml^{-1} streptomycin for 60 min at 37 °C, washed twice in prostate culture medium (DMEM supplemented with 10% FBS, 2 mM glutamine, 10 μg ml^{-1} insulin, 5.5 μg ml^{-1} transferrin, 6.7 ng ml^{-1} selenium, 1 nM testosterone (Innovative Research of America), 100 U ml^{-1} penicillin and 100 mg ml^{-1} streptomycin) and cultured in the same medium in 24-well plates coated with 10 μg ml^{-1} collagen type I. UGM cells were passaged at confluency by trypsin digestion and cultured *in vitro* for up to 1 week.

Prostate cell preparation. Human and mouse prostates were minced, placed in DMEM supplemented with 10% FBS, 2 mM glutamine, 100 U ml^{-1} penicillin and 100 mg ml^{-1} streptomycin, digested with 1 mg ml^{-1} collagenase/dispase for 90 min at 37 °C with agitation and passed through a 70 μm filter.

Serial isolation/transplantation *in vivo*. For secondary transplants of CD117$^+$ cells, primary grafts were magnetically sorted (~49,000 dissociated cells were obtained per primary graft, with CD117$^+$ cells constituting 19% of magnetically sorted cells), and sorted cells (10,000 cells per graft) were mixed with UGM stromal cells (250,000 cells per graft). For tertiary transplants of CD117$^+$ cells, secondary grafts were magnetically sorted (~40,000 dissociated cells were obtained per secondary graft, with CD117$^+$ cells constituting 11% of magnetically sorted cells) and sorted cells (2,200 cells per graft) were mixed with UGM stromal cells (250,000 cells per graft). For secondary transplants of Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ cells, primary grafts were sorted by FACS (~31,000 dissociated cells were obtained per secondary graft, with Lin$^-$Sca-1$^+$CD133$^+$CD44$^+$CD117$^+$ cells constituting 0.02% of viable FACS-sorted cells). Sorted cells (15 cells per graft) were mixed with UGM stromal cells (250,000 cells per graft). All serial transplantation grafts were collected 12 weeks after implantation. Gross graft images were acquired on a SMZ 800 dissecting microscope (Nikon) with a Coolpix 4300 digital camera (Nikon).

RNA isolation and qRT–PCR. Prostates from 8–10-week-old C57BL/6 mice were collected and teased apart to extend the tubules. For comparison of prostatic regions, each prostatic lobe was divided into distal, intermediate and proximal regions. For comparison of prostatic lobes, each prostate was divided into dorsal, lateral, ventral and anterior lobes. Total RNA was isolated using an RNeasy Mini kit (Qiagen) and qRT–PCR was performed with Power SYBR Green (Applied Biosystems) using the following primer sets: *Sca-1*, 5′-ATGGACACTTCTCACACTACAAAG-3′ and 5′-TCAGAGCAAGGTCTGCAGGAGGACTG-3′; *Cd44*, 5′-AATTCCGAGGATTCATCCCA-3′ and 5′-CGCTGCTGACATCGTCATC-3′; *Cd49b*, 5′-CCGGCATACGAAAGAATTGG-3′ and 5′-GAAGAGCTGAGGGTTATGT-3′; *Cd49f* (also known as *Itga6*), 5′-GTGGCCCAAGGAGATTAGC-3′ and 5′-GTTGACGCTGCAGTTGAGA-3′; *Cd133*, 5′-ACCAACACCAAGAACAAGGC-3′ and 5′-GGAGCTGACTTGAATTGAGG-3′; *Bcl2*, 5′-ATGTGTGTGGAGAGCGTCAAC-3′ and 5′-AGACAGCCAGGAGAAATCAAAC-3′; *Tert*, 5′-ATGGCGTTCCTGAGTATG-3′ and 5′-TTCAACCGCAAGACCGACAG-3′; *p63*, 5′-TTGTACCTGGAAAACAATG-3′ and 5′-TCGAAGCTGTGTGGGCCCGGG-3′; *Ck14*, 5′-GACTTCCGGACCAAGTTTGA-3′ and 5′-CTTGAGGCTCTCAATCTGC-3′; *Ck18*, 5′-ACTCCGCAAGGTGGTAGATG-3′ and 5′-GCCTCGATTTCTGTCTCCAG-3′; *Cd24*, 5′-TAAAGGACGCGTGAAAGGTTTGA-3′ and 5′-GACAAAATGGGTCTCCATTCCGCAC-3′; *Cd34*, 5′-ATGCAGGTCCACAGGGACACG-3′ and 5′-CTGTCCTGATAGATCAAGTAG-3′; *Cd117*, 5′-GACGCAACTTCCTTATGATC-3′ and 5′-TGGTTTGAGCATCTTCACGG-3′; *Slug*, 5′-TTTCTCCAGACCCTGGCTGCT-3′ and 5′-TTTTCCCCAGTGTGAGTTCTA-3′; *Gapdh*, 5′-ACTGGCATGGCCTTCCG-3′ and 5′-CAGGCGGCACGTCAGATC-3′. Gene expression was normalized to *Gapdh* using the ΔC_T method.

Flow cytometry. Prostate cells (non-lineage depleted) were permeabilized with 0.1% Triton X-100, stained with primary (CK14, CK18, APC-conjugated CD117) and secondary (Alexa Fluor 488 or 594) antibodies, and analysed on an LSR-II flow cytometer (Becton Dickinson).

Colony formation *in vitro*. The prostate colony formation assay has been described previously[13]. In brief, prostate cells from 8–10-week-old C57BL/6 mice were magnetically sorted into CD117$^{+/-}$ fractions, resuspended at 8,000 cells per 100 μl collagen type I at 3 mg ml^{-1} in DMEM, placed in flat-bottom 96-well plates for 1 h at 37 °C, and overlaid with prostate culture medium supplemented with 15 ng ml^{-1} epidermal growth factor (Roche). Medium was changed every 48 h and colony formation was assessed after 7 days.

Prostate culture *ex vivo*. Postnatal day 4 C57BL/6 mouse prostates were collected and placed on 8-μm pore-size cell culture inserts (BD Falcon), and inserts were placed into 24-well plates containing 300 μl DMEM/F-12 supplemented with 0.5% glucose, 2 mM glutamine, 10 μg ml^{-1} insulin, 5.5 μg ml^{-1} transferrin, 6.7 ng ml^{-1} selenium, 100 U ml^{-1} penicillin, 100 mg ml^{-1} streptomycin, and 25 ng ml^{-1} fungizone. Medium supplemented with function-blocking anti-CD117 antibody (25 μg ml^{-1}) was also used. Medium was changed and images were acquired every 48 h, and prostates were collected after 10 days. Images were acquired on a MZ16FA dissecting microscope (Leica) with a Retiga EXi digital camera (QImaging). Net growth in prostate area was quantified using MetaMorph software (Molecular Devices). Branch point quantification was performed on gross images of day 8 prostates.

Castration and androgen replacement. For microarray and qRT–PCR analysis, C57BL/6 mice at 8–10 weeks of age were used. On day 0, mice were castrated. On days 3 and 14 after castration, prostates from a subset of mice were collected. On day 14 after castration, testosterone pellets (15 mg per pellet per mouse) were implanted, and on day 17 (3 days after hormone replacement) prostates were collected. Total RNA was isolated using an RNeasy Mini kit and MOE430v2 Affymetrix chips were used for microarray analysis. To assess CD117 function during prostate regeneration *in vivo*, C57BL/6 mice at 8 weeks of age were used. Mice were castrated on day 0, and on day 12 after castration anti-ragweed control antibody (10 mg kg^{-1} in PBS) or function-blocking anti-CD117 antibody (10 mg kg^{-1} in PBS) were administered by intraperitoneal injection. On day 14 after castration testosterone pellets (15 mg per pellet per mouse) were implanted and on day 15 (1 day after hormone replacement) antibody treatments were administered. On day 19 (5 days after hormone replacement) prostates were collected, weighed and processed for histology. Prostate weights are expressed as the net increase compared to control prostates from day 14 after castration.

Immunohistochemistry. Optimal cutting temperature (OCT) compound-frozen tissues were sectioned at 8 μm, fixed in 4% paraformaldehyde (for CK14, CK18, CD117, Syp, Pbsn, β1 integrin, H-2k^b, CD31 and SMA) or methanol/acetone (1:1 (v/v); for E-cadherin, active caspase3, Ki67 and Nkx3-1), and incubated with primary antibody for 45 min and secondary antibody for 30 min. Human prostate specimens were fixed in formalin and sectioned at 6 μm, and antigen retrieval was performed with BD Retrievagen A (BD Biosciences). For specificity controls, species-matched primary antibodies were used. Images were acquired on an Axioplan 2 imaging microscope (Zeiss) with an ORCA-ER digital camera (Hamamatsu). For *ex vivo* prostates, percentages of positive cells were quantified by assessing at least 600 cells for CK14 or CK18, 400 cells for Ki67 and 1,200 cells for E-cadherin. For regenerated *in vivo* prostates, the percentages of positive cells were quantified by assessing at least 800 cells for CK14 or CK18 and 2,500 cells for Ki67.

Laser capture microdissection and PCR-based genotyping. Single-cell-derived grafts were sectioned at 8 μm, mounted on metal frame membrane slides (Molecular Machines & Industries) and stained with mouse-specific β1 integrin and CD31. Within the graft, β1 integrin$^+$CD31$^-$ cells were isolated with a Nikon E2000 CellCut laser capture microdissector (Molecular Machines & Industries).

For *Foxn1*$^{+/+}$ cell controls, rat stromal cells (β1 integrin^{-}CD31^{-}) within the graft were isolated. For *Foxn1*$^{+/-}$ cell controls, athymic nu/nu kidney cells (β1 integrin^{+}) adjacent to the graft were isolated. Captured cells were lysed with a PicoPure DNA Extraction Kit (Molecular Devices) and PCR-based genotyping was performed (http://jaxmice.jax.org/pub-cgi/protocols/protocols.sh). Genomic DNA was amplified by PCR with primers for *Foxn1* (5'-GGCCCAGCAGGCAGCCCAAG-3′ and 5′-AGGGATCTCCTCAAAGGCT-TC-3′), digested with BsaJI and run on a 4% agarose gel. The undigested PCR product is 168 base pairs (bp); the digested *Foxn1*$^{+/+}$ PCR product gives 90-, 58- and 20-bp fragments; the digested *Foxn1*$^{+/-}$ PCR product gives 110-, 90-, 58- and 20-bp fragments. The absence of a 110-bp product indicates that genomic DNA is derived from *Foxn1*$^{+/+}$ (wild type) cells. PCR control reactions included water (negative control) and wild-type mouse genomic DNA (positive control).

Confocal and single cell microscopy. Confocal images were scanned and acquired with a LSM 510 META confocal microscope (Zeiss). Single cell images were acquired on an Eclipse TE300 inverted microscope (Nikon) with a Cascade Photometrics digital camera (Roper Scientific).

Limiting dilution analysis. Limiting dilution analysis was performed using the 'limdil' function in the 'statmod' software package (http://bioinf.wehi.edu.au/software/limdil/index.html). A confidence interval of 95% was used.

Statistical analysis. Group differences were evaluated using a two-tailed Student's *t*-test. *P* values of less than 0.05 were considered significant.

Vol 456 | 11 December 2008 | doi:10.1038/nature07424

LETTERS

A role for VEGF as a negative regulator of pericyte function and vessel maturation

Joshua I. Greenberg[1], David J. Shields[2], Samuel G. Barillas[1], Lisette M. Acevedo[2], Eric Murphy[2], Jianhua Huang[2], Lea Scheppke[2], Christian Stockmann[3], Randall S. Johnson[3], Niren Angle[1] & David A. Cheresh[2]

Angiogenesis does not only depend on endothelial cell invasion and proliferation: it also requires pericyte coverage of vascular sprouts for vessel stabilization[1,2]. These processes are coordinated by vascular endothelial growth factor (VEGF) and platelet-derived growth factor (PDGF) through their cognate receptors on endothelial cells and vascular smooth muscle cells (VSMCs), respectively[3,4]. PDGF induces neovascularization by priming VSMCs/pericytes to release pro-angiogenic mediators[5–7]. Although VEGF directly stimulates endothelial cell proliferation and migration, its role in pericyte biology is less clear. Here we define a role for VEGF as an inhibitor of neovascularization on the basis of its capacity to disrupt VSMC function. Specifically, under conditions of PDGF-mediated angiogenesis, VEGF ablates pericyte coverage of nascent vascular sprouts, leading to vessel destabilization. At the molecular level, VEGF-mediated activation of VEGF-R2 suppresses PDGF-Rβ signalling in VSMCs through the assembly of a previously undescribed receptor complex consisting of PDGF-Rβ and VEGF-R2. Inhibition of VEGF-R2 not only prevents assembly of this receptor complex but also restores angiogenesis in tissues exposed to both VEGF and PDGF. Finally, genetic deletion of tumour cell VEGF disrupts PDGF-Rβ/VEGF-R2 complex formation and increases tumour vessel maturation. These findings underscore the importance of VSMCs/pericytes in neovascularization[8,9] and reveal a dichotomous role for VEGF and VEGF-R2 signalling as both a promoter of endothelial cell function and a negative regulator of VSMCs and vessel maturation.

VEGF and PDGF contribute to angiogenesis and vessel stabilization by activating endothelial cells and VSMCs respectively, suggesting that a combination of these factors might elicit a synergistic response. To test this directly in angiogenesis models, PDGF, VEGF or both were mixed with Matrigel and implanted subcutaneously into mice or topically applied to the chorioallantoic membrane (CAM) of 10-day-old chick embryos. As expected, PDGF-BB or VEGF, at an optimal angiogenic dose (Supplementary Fig. 1a, b), promotes robust neovascularization (Fig. 1a). Surprisingly, the

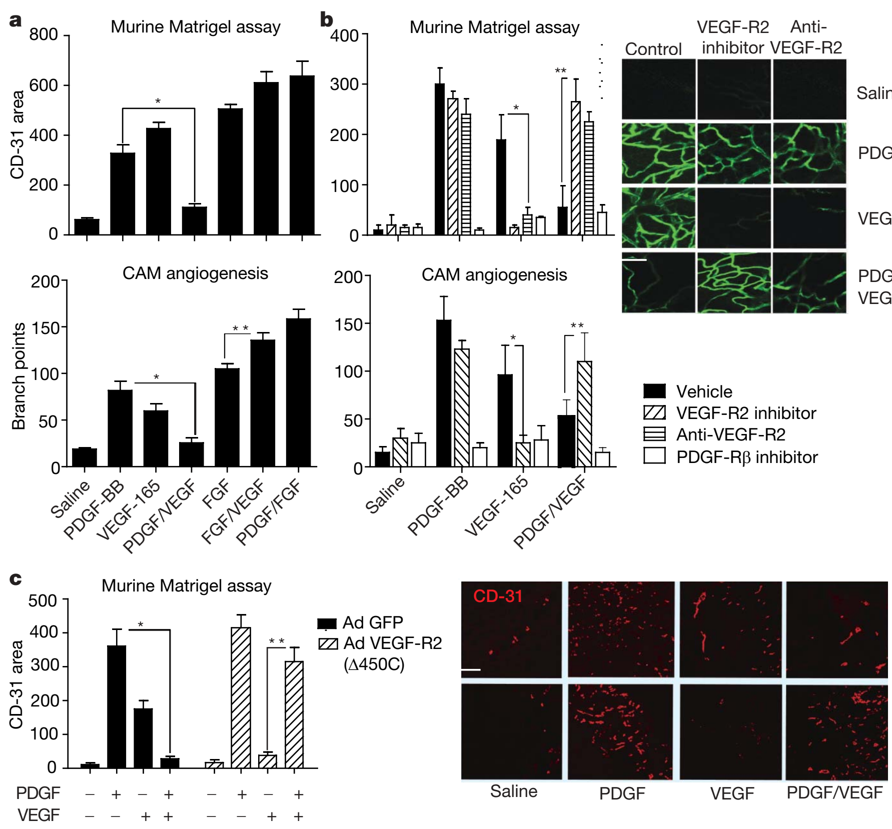


Figure 1 | VEGF inhibits PDGF-mediated angiogenesis through VEGF-R2. a, Upper: vessel growth into Matrigel as measured by CD-31$^+$ area; $*P < 0.02$; lower: branch points quantified on filter paper impregnated with saline or growth factors placed on embryonic day-10 chick CAMs; $*P < 0.02$, $**P < 0.05$, results from one-way analysis of variance (ANOVA). Further images are displayed in Supplementary Fig. 1d. **b**, Upper: vessel growth into Matrigel as in **a**, with the application of inhibitors; $*P < 0.001$, $**P < 0.001$; lower: branch points quantified on filter paper impregnated with saline or growth factor as in **a**, with the application of inhibitors; $*P < 0.02$, results from two-way ANOVA, $**P < 0.05$; right: confocal microscopy of Matrigel from mice injected intravenously with fluorescent *Griffonia* lectin; scale bar, 100 μm. **c**, CD-31+ area quantified by image analysis of Matrigel (right panel) impregnated with saline or growth factor implanted into mice infected with adenovirus expressing green fluorescent protein (GFP) only or GFP and truncated VEGF-R2(Δ450C); scale bar, 200 μm; $*P < 0.001$, $**P < 0.001$, results from one-way ANOVA. All error bars, s.d.; $n > 5$ per group in all panels.

[1]Department of Surgery, School of Medicine, [2]Department of Pathology and Moore's UCSD Cancer Center, [3]Section of Molecular Biology, Division of Biology, University of California, San Diego, California 92093, USA.

combination of these cytokines results in complete suppression of angiogenesis (Fig. 1a). Inhibition of angiogenesis is not likely due to excessive proliferation of mural or stromal cells, because PDGF and VEGF each combine with bFGF to produce an increased angiogenic response (Fig. 1a)[10]. Thus, a specific antagonistic relationship exists between PDGF and VEGF during neovascularization.

We next considered whether PDGF or VEGF receptors contribute to the antagonistic effects of combined VEGF/PDGF treatment on neovascularization. After PDGF receptor blockade, we observed suppression of angiogenesis in the presence of either growth factor (Fig. 1b), consistent with a role for PDGF in the release of pro-angiogenic factors and the recruitment of VSMCs/pericytes to nascent endothelial cell sprouts[11,12]. Although combined PDGF/VEGF stimulation leads to the appearance of tenuous and small-calibre vascular sprouts on CAMs at 48 h, a complete absence of new vessel growth is evident on CAMs by 72 h (Fig. 1b and Supplementary Fig. 1c). Thus, combined PDGF/VEGF treatment initiates vessel sprouting but not a durable vascular response.

To assess the role of VEGF-R2 as a negative regulator of pericyte function, Matrigel and CAM tissues were treated with either a VEGF-R2 kinase inhibitor or anti-VEGF-R2 and stimulated with PDGF, VEGF or both. As expected, either inhibitor disrupts VEGF- but not PDGF-mediated angiogenesis in both models (Fig. 1b)[13]. Interestingly, after PDGF/VEGF treatment, VEGF-R2 inhibition restores angiogenesis to a level achieved with either growth factor alone (Fig. 1b), suggesting that under conditions of VEGF/PDGF stimulation, VEGF-R2 negatively regulates neovascularization. To validate these observations, mice were injected with Matrigel impregnated with growth factors, as well as adenovirus expressing an inactive truncation mutant of the VEGF-R2 receptor. Gene delivery to the vascular compartment suppresses VEGF-R2 phosphorylation in these tissues (Supplementary Fig. 2a, b). Importantly, the VEGF-R2 truncation mutant restores the angiogenic response after PDGF/VEGF stimulation without affecting PDGF-mediated angiogenesis (Fig. 1c). These studies demonstrate that, in the presence of both VEGF and PDGF, VEGF-R2 can limit pro-angiogenic events.

We next considered the effect of VEGF treatment on PDGF-stimulated pericyte recruitment to nascent vessels during angiogenesis. As expected, PDGF treatment alone produces a robust influx of alpha-smooth-muscle actin-expressing (α-SMA+; VSMC/pericyte marker) cells to the surface of patent blood vessels (Fig. 2a)[5]. However, stimulation of tissues with PDGF/VEGF results in a reduction in vascular pericyte coverage in comparison with tissues treated with PDGF alone. Importantly, pericyte coverage and neovascularization are restored after VEGF-R2 inhibition (Fig. 2a), suggesting that VSMC/pericyte VEGF-R2 negatively regulates this response.

To assess whether VSMC function could be influenced by the combination of PDGF and VEGF, we examined the effects of these growth factors on proliferation and migration in primary human VSMCs *in vitro*. Although PDGF induces VSMC proliferation and migration, exposure to both PDGF and VEGF results in a complete suppression of these responses (Fig. 2b). In contrast, human umbilical vein endothelial cells (HUVECs), which express VEGF-R2 but not PDGF receptors, and normal human dermal fibroblasts (NHDFs), which express PDGF but not VEGF receptors (Supplementary Fig. 3a), exhibit robust migration and proliferation in the presence of PDGF/VEGF. Similar results were obtained with a wide range of other endothelial cells and VSMC/pericytes (Supplementary Fig. 4). Importantly, VSMCs pre-treated with a VEGF-R2 inhibitor or subjected to short interfering RNA (siRNA)-mediated knockdown of VEGF-R2, but not VEGF-R1, demonstrate complete reversal of PDGF/VEGF-mediated suppression of cell migration and proliferation (Fig. 2c). This may be explained by the finding that VEGF significantly attenuates PDGF-induced pp60-Src and ERK signalling (Supplementary Fig. 5). Importantly, co-stimulation with bFGF and PDGF does not suppress PDGF-mediated VSMC chemotaxis (data not shown), further supporting a specific relationship between VEGF and PDGF.

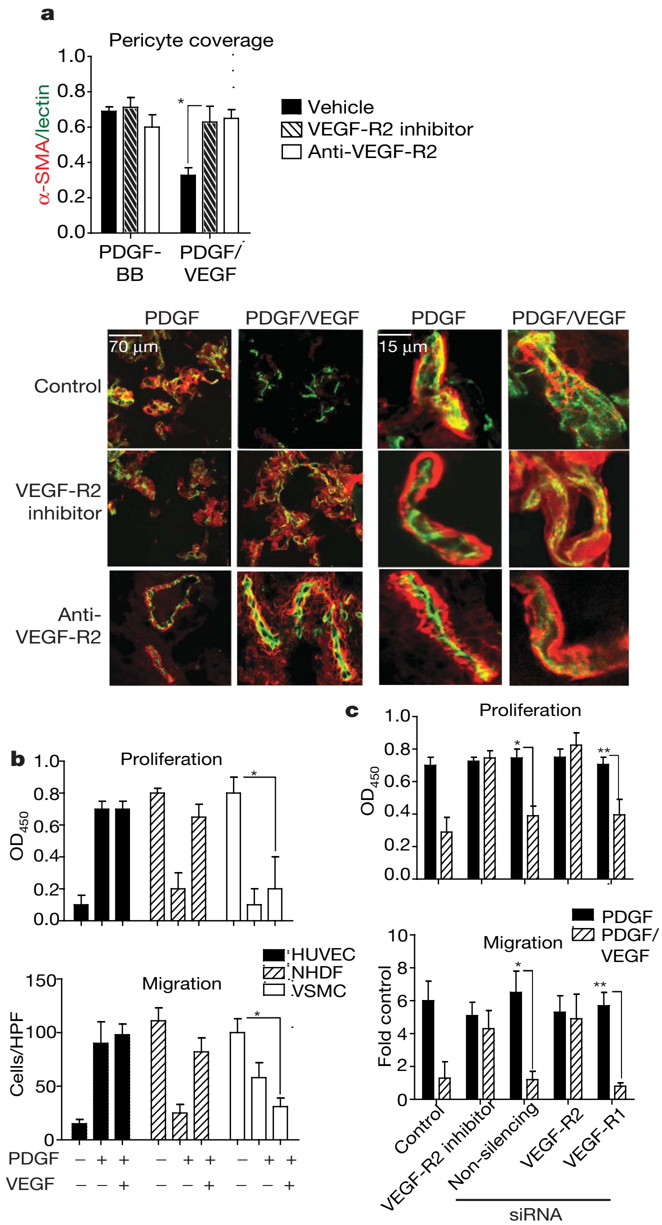


Figure 2 | VEGF disrupts pericyte coverage/VSMC activation through VEGF-R2. **a**, Bottom: confocal microscopy images of reconstituted basement membrane impregnated with growth factors from mice perfused with fluorescent *Griffonia* lectin (green), stained for α-SMA (red); top: pericyte coverage as assessed by α-SMA/lectin co-localization; $*P < 0.01$, results from one-way ANOVA, $n = 6$ animals per group. **b**, Representative proliferation (upper, $*P < 0.001$) and migration assays as measured by cells per high-powered field (HPF) (lower, $*P < 0.001$); results from one-way ANOVA. **c**, Representative VSMC proliferation (upper, $*P < 0.01$, $**P < 0.05$) and migration (lower, $*P < 0.001$, $**P < 0.001$) assays performed after siRNA-mediated knockdown of VEGF receptors normalized to PBS control where indicated; results from one-way ANOVA, $n > 3$ experiments. All error bars, s.d.

The antagonistic results of PDGF/VEGF treatment do not result from competition for receptor occupancy because treatment with a range of VEGF doses fails to suppress PDGF stimulation of PDGF-Rβ in NHDFs, which lack VEGF-R2 (data not shown). Thus, it appears that VEGF-ligation of VEGF-R2 in VSMCs leads to inactivation of PDGF-Rβ, the dominant PDGF receptor on these cells. To test this possibility, VSMCs were exposed to varying concentrations of VEGF (in the presence or absence of PDGF) and the PDGF-Rβ phosphorylation (activation) status was evaluated. Although VEGF alone fails

to induce PDGF-Rβ phosphorylation, it suppresses PDGF-mediated PDGF-Rβ tyrosine phosphorylation fivefold, which can be reversed by PDGF titration (Supplementary Fig. 6a, b). Importantly, VEGF-121, which lacks a heparin-binding domain, also suppresses PDGF-mediated activation of PDGF-Rβ (Supplementary Fig. 6c). To evaluate the requirement of VEGF receptors for this process, we subjected VSMCs to VEGF-R1 or VEGF-R2 siRNA-mediated knockdown and analysed the effect of PDGF/VEGF co-stimulation on PDGF-Rβ activity. Knockdown of VEGF-R2 but not VEGF-R1 abrogates the VEGF-mediated inactivation of PDGF-Rβ, indicating that VEGF-R2 is essential for VEGF-mediated suppression of PDGF signalling in VSMCs (Fig. 3a). VEGF-R2 kinase inhibition also restores PDGF-Rβ phosphorylation in cells exposed to PDGF/VEGF, providing further evidence that VEGF-R2 negatively regulates PDGF-Rβ (Supplementary Fig. 7a). Thus, co-stimulation of VSMCs with PDGF/VEGF results in cross-talk between the cognate receptors for these ligands, leading to suppression of PDGF signalling.

To define the underlying mechanism of VEGF-R2-dependent inhibition of PDGF-Rβ activity further, VEGF-R2 or VEGF-R2(Δ450C) were expressed in NHDF cells, which lack VEGF-R2. VEGF induces phosphorylation of wild-type VEGF-R2 but not the truncated derivative (Supplementary Fig. 8a, b). NHDF cells transduced with VEGF-R2 or VEGF-R2(Δ450C) demonstrate equivalent phosphorylation of PDGF-Rβ in response to PDGF alone (Fig. 3b). However, when NHDFs expressing VEGF-R2 were stimulated with VEGF/PDGF, PDGF-Rβ exhibited decreased phosphorylation relative to cells expressing truncated VEGF-R2. This was associated with a concomitant reduction in proliferation and cell migration, relative to cells exposed to either cytokine alone, demonstrating that VEGF-R2 expression is sufficient to inhibit PDGF-dependent function in VSMCs (Fig. 3b).

Given the functional crosstalk between VEGF-R2 and PDGF-Rβ in VSMCs, we evaluated whether these receptors interact directly by performing reciprocal co-immunoprecipitation studies. Importantly, treatment of VSMCs with PDGF/VEGF results in the formation of a previously undescribed VEGF-R2/PDGF-Rβ complex, whereas no interaction is detected after exposure of cells to either growth factor alone (Fig. 3c). The same biochemical interaction was not detected with PDGF-Rα (Supplementary Fig. 9a). Furthermore, pre-treatment of VSMCs with a VEGF-R2 inhibitor attenuates VEGF-R2/PDGF-Rβ complex formation, suggesting a requirement for VEGF-R2 activity. Similar findings are observed in the pericyte-like 10T1/2 cell line, which is often used to approximate pericytes *in vitro*, and which has been found to exhibit growth suppression in co-culture with endothelial cells (Supplementary Fig. 9b)[14,15]. As a final biochemical confirmation, growth-factor-dependent PDGF-Rβ/VEGF-R2 complex formation was identified in HEK-293 cells ectopically expressing both VEGF-R2 and PDGF-Rβ (Supplementary Fig. 9c). To confirm these findings by using a cell biological approach, we performed a proximity ligation assay (PLA), which depicted close proximity of cellular molecules and demonstrated a fivefold increase in the VEGF-R2/PDGF-Rβ interaction after PDGF/VEGF co-stimulation, relative to cells stimulated with either growth factor alone (Supplementary Fig. 10). To characterize this interaction further, NHDFs were transduced with full-length or VEGF-R2(Δ450C). As demonstrated in VSMCs, expression of full-length VEGF-R2 but not the truncation mutant results in the formation of a VEGR-R2/PDGF-Rβ complex (after dual cytokine stimulation only) (Fig. 3d). These results demonstrate a requirement

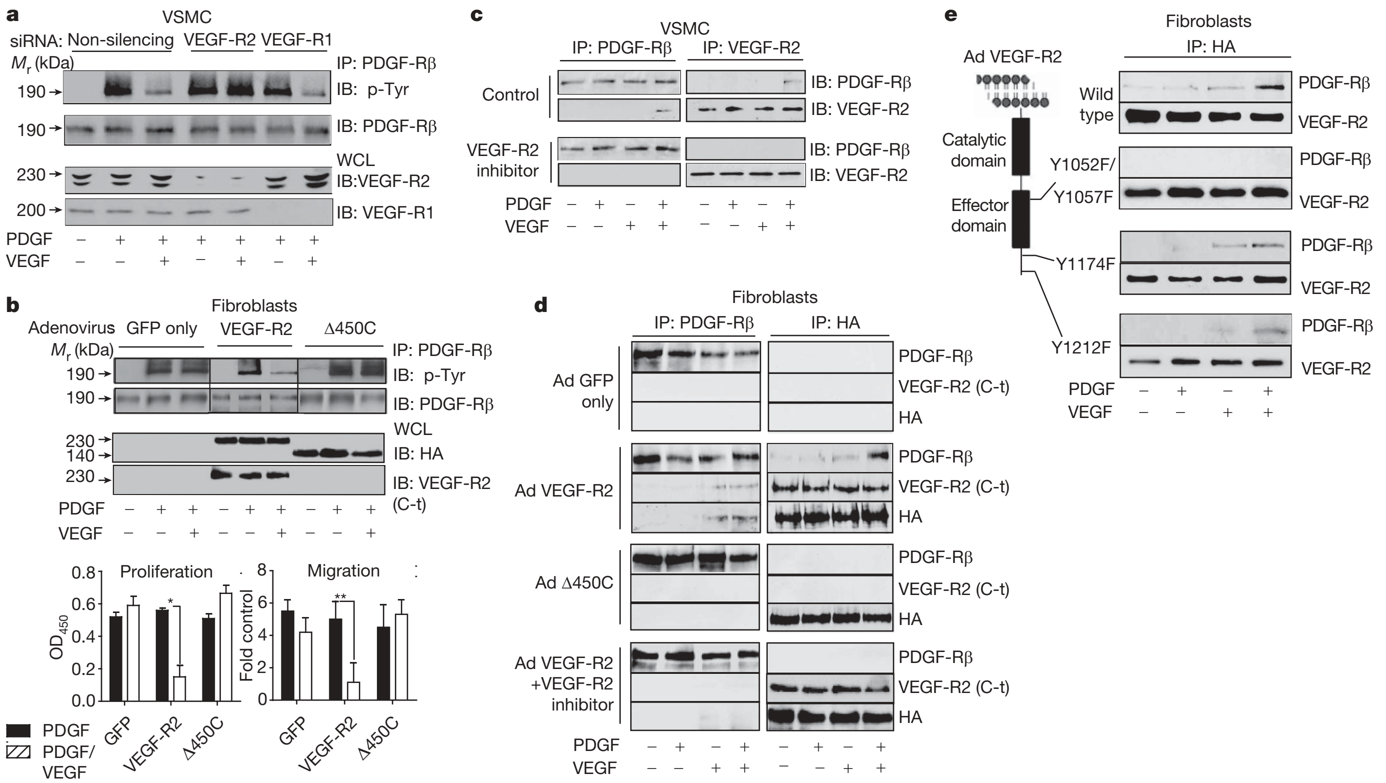


Figure 3 | An inducible VEGF-R2/PDGF-Rβ complex forms in VSMC. **a**, Immunoblot analysis of anti-PDGF-Rβ immunoprecipitation lysates from serum-starved VSMCs subjected to siRNA-mediated VEGF receptor knockdown. **b**, Upper: immunoblot analysis of anti-PDGF-Rβ immunoprecipitated lysates from serum-starved NHDFs (VEGF-R2 null) infected with adenovirus. Immunoblots represent at least three similar experiments. Lower: representative proliferation (left) and migration (right) experiments with VEGF-R2-reconstituted NHDFs; $*P < 0.001$, $**P < 0.02$, results from one-way ANOVA, $n > 3$ separate experiments. Error bars, s.d. **c**, **d**, **e**, Representative reciprocal co-immunoprecipitation followed by immunoblot analysis with lysates derived from serum-starved VSMCs pre-treated with either control or VEGF-R2 inhibitor (**c**), serum-starved NHDFs transduced with truncated VEGF-R2 adenovirus (**d**) or NHDFs transduced with VEGF-R2 catalytic site (Y1052F and Y1057F) and effector site mutants (Y1174F and Y1212F) (**e**). IB, immunoblot; IP, immunoprecipitation; WCL, whole-cell lysate; C-t, C terminus. Immunoblots are representative of at least three similar experiments; mutant receptors are characterized in Supplementary Fig. 8.

for the VEGF-R2 carboxy (C) terminus in the VEGR-R2/PDGF-Rβ interaction.

We next transduced NHDFs with VEGF-R2(Y1052/1057F), which fails to undergo auto-phosphorylation in response to VEGF (Supplementary Fig. 8e). Similar to the results observed after transduction of VEGF-R2(Δ450C), this catalytically inactive mutant fails to complex with PDGF-Rβ after treatment with dual growth factor, demonstrating that VEGF-R2 activity is required for interaction with PDGF-Rβ (Fig. 3e). Importantly, complex formation is still observed when two VEGF-R2 effector sites (Y1174 and Y1212) are mutated.

Tumours express high levels of VEGF and as a result develop tortuous, leaky and immature blood vessels[16,17]. Anti-VEGF therapy results in vessel normalization characterized by increased pericyte coverage, tumour perfusion and chemotherapeutic sensitivity[18]. To investigate the role of VEGF in tumour vessel normalization, we examined the vasculature of fibrosarcomas that either express or lack VEGF[19]. Similar to results from anti-VEGF therapy, $VEGF^{-/-}$ tumours produce mature vessels with extensive pericyte coverage, whereas wild-type ($VEGF^{+/+}$) tumours develop an immature tumour vasculature with minimal pericyte coverage (Fig. 4a, b). Importantly, dual expression of PDGF-Rβ and VEGF-R2 appears limited to α-SMA+ perivascular cells (Supplementary Fig. 11b). Interestingly, the VEGF-R2/PDGF-Rβ complex is only detected in lysates of $VEGF^{+/+}$ tumours, which feature an immature vasculature (Fig. 4c). By contrast, complex formation is attenuated in $VEGF^{-/-}$ tumours, concomitant with increased PDGF-Rβ phosphorylation (Fig. 4c) and decreased VEGF-R2 phosphorylation (Supplementary Fig. 11a). Next, we implanted human pancreatic tumour cells orthotopically into the tail of the mouse pancreas and allowed tumours to develop for two weeks. Analysis of immunoprecipitated protein from these tumours demonstrated PDGF-Rβ/VEGF-R2 complex formation (Supplementary Fig. 12a). Although the tumour cells within these tissues express negligible levels of VEGF or PDGF receptors, we did observe co-expression of VEGF-R2 and PDGF-Rβ on perivascular cells that co-stain with α-SMA (Supplementary Fig. 12b, c). Taken together, these studies demonstrate that VEGF can induce the formation of a VEGF-R2/PDGF-Rβ complex implicated both in the suppression of pericyte function and in the regulation of tumour angiogenesis. Our findings support a mechanism where VEGF inhibits pericyte function and blood vessel maturation through the induction of a VEGF-R2/PDGF-Rβ complex, and may in part, explain the molecular basis of anti-VEGF therapy in cancer patients.

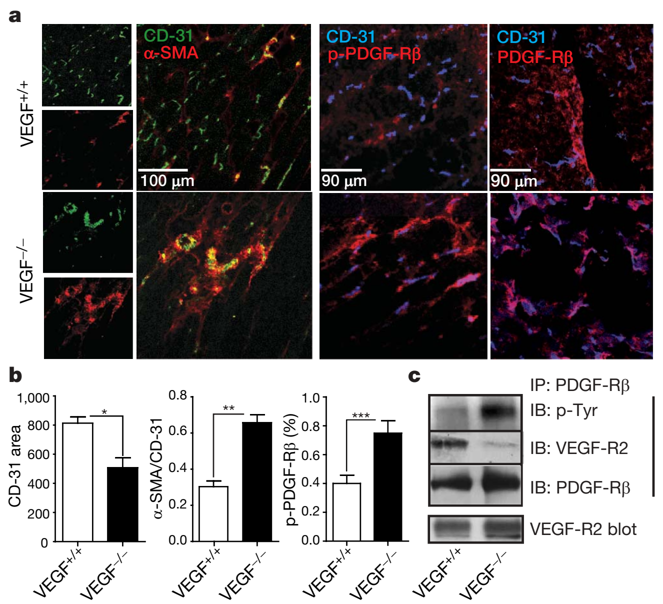


Figure 4 | VEGF loss attenuates VEGF-R2/PDGF-Rβ complex formation and tumour vessel maturation. a, Confocal microscopy of wild-type ($VEGF^{+/+}$) and knockout ($VEGF^{-/-}$) fibrosarcomas demonstrating histological markers of vascular maturity and phosphorylated PDGF-Rβ (Y751); **b**, image analysis: $P < *0.01$, $**P < 0.001$, $***P < 0.05$, results from unpaired t-test, $n > 3$ animals per group. Error bars, s.d. **c**, Anti-PDGF-Rβ immunoprecipitated tumour lysates derived from $VEGF^{+/+}$ and $VEGF^{-/-}$ tumours.

METHODS SUMMARY

HUVECs, HMVECs, HAECs and NHDFs were acquired from Clonetics. Human aortic VSMCs were isolated from three donors (Clonetics and Cascade Biologicals). The pericyte line C3 10T1/2 (clone 8) was acquired from the American Type Culture Collection (ATCC). Murine lung endothelial cells (MLECs) were isolated by immunoselection using Dynal magnetic beads (Invitrogen). For *in vitro* studies, unless otherwise indicated, PDGF-BB, VEGF-165 and FGF-2 were used at 20 ng ml^{-1}, 40 ng ml^{-1} and 50 ng ml^{-1}, respectively. Growth factor (500 and 100 ng, respectively) was used for Matrigel and CAM experiments. Epitope-tagged mutant receptor constructs and adenoviruses were generated using standard cloning techniques and validated by DNA sequencing. The procedures for CAM[20], Matrigel[21] and murine fibrosarcoma[19] experiments were performed with institutional approval as previously reported or with slight modifications.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 March; accepted 10 September 2008.
Published online 9 November 2008.

1. Carmeliet, P. Angiogenesis in life, disease and medicine. *Nature* **438**, 932–936 (2005).
2. Andrae, J. *et al.* Role of platelet-derived growth factors in physiology and medicine. *Genes Dev.* **22**, 1276–1312 (2008).
3. Ferrara, N., Gerber, H. P. & LeCouter, J. The Biology of VEGF and its receptors. *Nature Med.* **9**, 669–676 (2003).
4. Inui, H. *et al.* Differences in signal transduction between platelet-derived growth factor (PDGF) alpha and beta receptors in vascular smooth muscle cells. PDGF-BB is a potent mitogen, but PDGF-AA promotes only protein synthesis without activation of DNA synthesis. *J. Biol. Chem.* **269**, 30546–30552 (1994).
5. Vincent, L. & Rafii, S. Vascular frontiers without borders: multifaceted roles of platelet-derived growth factor (PDGF) in supporting postnatal angiogenesis and lymphangiogenesis. *Cancer Cell* **6**, 307–309 (2004).
6. Pietras, K. *et al.* Functions of paracrine PDGF signaling in the proangiogenic tumor stroma revealed by pharmacological targeting. *PLoS Med.* **5**, e19 (2008).
7. Jechlinger, M. *et al.* Autocrine PDGFR signaling promotes mammary cancer metastases. *J. Clin. Invest.* **116**, 1561–1570 (2006).
8. Song, S. *et al.* PDGFRβ+ perivascular progenitor cells in tumours regulate pericyte differentiation and vascular survival. *Nature Cell Biol.* **7**, 870–879 (2005).
9. Lindahl, P., Johansson, B. R., Leveen, P. & Betsholtz, C. Pericyte loss and microaneurysm formation in PDGF-B-deficient mice. *Science* **11**, 242–245 (1997).
10. Nissen, L. J. *et al.* Angiogenic factors FGF-2 and PDGF-BB synergistically promote murine tumor neovascularization and metastasis. *J. Clin. Invest.* **117**, 2766–2777 (2007).
11. Abramsson, A., Lindblom, M. & Betsholtz, C. Endothelial and nonendothelial sources of PDGF-B regulate pericyte recruitment and influence vascular pattern formation in tumors. *J. Clin. Invest.* **112**, 1142–1151 (2003).
12. Sennino, B. *et al.* Sequential loss of tumor vessel pericytes and endothelial cells after inhibition of PDGF-B by selective aptamer AX102. *Cancer Res.* **67**, 7358–7367 (2007).
13. Kim, K. J. *et al.* Inhibition of vascular endothelial growth factor-induced angiogenesis suppresses tumour growth *in vivo. Nature* **362**, 841–844 (1993).
14. Hirschi, K. K. *et al.* Endothelial cells modulate the proliferation of mural cell precursors via platelet-derived growth factor-BB and heterotypic cell contact. *Circ. Res.* **84**, 298–305 (1999).
15. Hirschi, K. K. *et al.* PDGF, TGF-β, and heterotypic cell-cell interactions mediate endothelial cell-induced recruitment of 10T1/2 cells and their differentiation to a smooth muscle fate. *J. Cell Biol.* **141**, 805–814 (1998).
16. Morikawa, S. *et al.* Abnormalities in pericytes on blood vessels and endothelial sprouts in tumors. *Am. J. Pathol.* **160**, 985–1000 (2002).
17. Jain, R. K. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science* **307**, 58–62 (2005).
18. Willett, C. G. *et al.* Direct evidence that the VEGF-specific antibody bevacizumab has antivascular effects in human rectal cancer. *Nature Med.* **2**, 145–147 (2004).
19. Grunstein, J. *et al.* Tumor-derived expression of vascular endothelial growth factor is a critical factor in tumor expansion and vascular function. *Cancer Res.* **23**, 2800–2810 (2004).
20. Eliceiri, B. *et al.* Selective requirement for Src kinases during VEGF-induced angiogenesis and vascular permeability. *Mol. Cell* **4**, 915–924 (1999).

21. Alavi, A. *et al.* Chemoresistance of endothelial cells induced by basic fibroblast growth factor depends on Raf-1-mediated inhibition of the proapoptotic kinase, ASK1. *Cancer Res.* **6**, 2766–2772 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Stupack and D. Schlaepfer for advice and review of the manuscript. We also thank A. Tomar for assistance with the PLA imaging; C. Heldin for the PDGF-Rβ complementary DNA; and J. Desgrosellier for processing pancreatic tumour samples. This work was supported by National Institutes of Health grants R37-CA50286 (D.A.C.) and CA082515 (R.S.J.), and by a University of California Academic Senate Grant (N.A.). C.S. was funded by the Deutsche Forschungsgemeinschaft (DFG STO 787/1-1), and L.M.A. received an Institutional Research and Academic Career Development Award postdoctoral fellowship (National Institutes of Health grant GM 68524).

Author Contributions J.I.G. generated mutant receptors, performed biochemical and cell imaging analyses, and prepared the manuscript. D.J.S. and E.M. assisted with experiments and manuscript preparation. S.G.B. performed angiogenesis assays. L.M.A. assisted with adenovirus production, cell isolation and tumour staining. J.H. and L.S. performed RNA analysis and assisted with biochemical analyses. C.S. and R.S.J. generated, implanted and harvested fibrosarcomas. D.A.C. and N.A. supervised and directed the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.A.C. (dcheresh@ucsd.edu).

METHODS

Cell culture and reagents. PDGF-BB (referred to as PDGF in the text) and VEGF-165 (referred to as VEGF in the text) were acquired from Peprotech; FGF-2 was from Chemicon. Endothelial cells were grown in EBM-2 media, NHDFs were grown in FGM media (Clonetics) and VSMCs were grown in Medium-231 (Cascade Biologicals). HEK-293T (ATCC) and C3 10T1/2 cells were grown in DMEM with 15% FBS-supplemented penicillin/streptomycin. Inhibitors SU-5416 (ref. 22) (VEGF-R2) and AG-1296 (PDGF-R) used at 1 μM were purchased from Calbiochem. Anti-VEGF-R2 (MAB-4431, R&D) was used at 50 μg ml^{-1} in Matrigel experiments.

Antibodies to PDGF-Rα (sc-338), PDGF-Rβ (sc-432), VEGF-R2 (sc-504), VEGF-R1 (sc-9029), ERK-2 (sc-1647), FAK (sc-557) and HRP-conjugated PY-20 (sc-508) were purchased from Santa Cruz Biotechnology. Other antibodies included p-ERK (Cell Signalling), p-FAK (Y861, Biosource), p-VEGF-R2 (Y1212, Upstate), p-VEGF-R2 (Y1052, Upstate) and p-VEGF-R2 (Y1175, Cell Signalling).

Co-immunoprecipitation/immunoblotting. Two million cells in 150-cm dishes were starved in Basal SMC Media (Medium 231) with 0.2% FBS for 18 h. addition of growth factor was followed by a chilled PBS wash, and cell lysis proceeded after seven minutes in RIPA buffer (5 mM Tris pH 7.4, 150 mM NaCl, 1% Triton, 0.1 mM NaF) containing protease inhibitors (Minitab, Roche), 1 mM NaOV, and 10 mM PMSF. Clarified lysates were pre-incubated for 1 h with antibody followed by the addition of protein A/G beads (Pierce) for end-over-end tumbling at 4 °C overnight. Equal amounts of protein were subsequently resolved by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and immunoblotting followed by detection with SuperSignal ECL (Pierce). Immunoprecipitation and immunoblotting of PDGF-Rβ was with anti-PDGF-Rβ (sc-432) and anti-PDGF-Rβ (sc-339), respectively. Immunoprecipitation and immunoblotting of human VEGF-R2 was with anti-VEGF-R2 (sc-19530) and anti-VEGF-R2 (sc-504), respectively. Immunoprecipitation and immunoblotting of murine VEGF-R2 was with anti-VEGF-R2 (sc-504) and VEGF-R2 (sc-6251), respectively.

Chemotaxis (Transwell) assay. Transwell assays were performed as previously described[23].

Proliferation (XTT) assay. Five thousand cells per well were seeded in 96-well plates. Cells were starved in Basal SMC Media (Medium 231) with 0.2% FBS for 18 h before the addition of growth factors. Incubation was for 48 h before addition of tetrazolium (XTT) salt from a Tox2 Kit (Sigma).

VEGF receptor siRNA knockdown experiments. One million VSMCs were electroporated (Amaxa) with 80 pmol of scramble or VEGF-R1 or VEGF-R2 siRNA duplexes (Santa Cruz Biotechnology) 48 h before starvation for growth factor treatment experiments according to the protocol for AoSMC Nucleofection (Amaxa). Cells were recovered in 100-mm dishes in complete media. The following pooled siRNA sequences (Santa Cruz Biotechnology) were used to knockdown VEGF-R2 and VEGF-R1 respectively:

VEGF-R2 siRNA sequences (mRNA accession number, NM_002253): duplex 1 sense strand, ACUGUGGUGAUUCCAUGUCTT; duplex 2 sense strand, ACUUGUAAACCGAGACCUATT; duplex 3 sense strand, CACCUGUUUGCAAGAACUUTT; duplex 4 sense strand, AGACUACGUUGGAGCAAUCTT. VEGF-R1 siRNA sequences (mRNA accession number, NM_002019): duplex 1 sense strand, CACAUCGACAAACCAAUACTT; duplex 2 sense strand, ACACGAGAGUUCAAAUGACTT; duplex 3 sense strand, CAGAGACGUUACUUGGAUUTT; duplex 4 sense strand, CAACAGGAUGGUAAAGACUTT.

VEGF receptor adenovirus gene transfer. High-titre adenovirus was generated with the Adeasy system[24] (Stratagene) and applied to cells at a multiplicity of infection (MOI) of 50 plaque-forming units 48 h before experiments. Twenty microlitres of 10^{11} plaque-forming units per millilitre titre viral stock was applied with growth factor for Matrigel assays.

Receptor transient transfection experiments. Two million HEK-293 cells seeded into 150-cm dishes were co-transfected with 10 μg pcDNA3 (Invitrogen) containing VEGF-R2 (cloned as a *Not*I/*Xho*I fragment from pBluescript SK- described below) or PDGF-Rβ (a gift from C. Heldin) cDNAs using Lipofectamine 2000 (Invitrogen). Cells were starved 48 h later for growth-factor treatments.

VEGF-R2 cloning and mutagenesis. Full-length murine VEGF-R2 cDNA was generated by PCR from a template reverse-transcribed from murine lung RNA (see RT–PCR) using Platinum Taq High Fidelity (Invitrogen) with a DNA Engine (PTC-200, Bio-Rad cycler as follows: 94 °C for 30 s, followed by 94 °C for 15 s, 55 °C for 30 s, and 68 °C for 5 min for 30 cycles). *Not*I and *Xho*I restriction sites (underlined) were introduced by PCR to generate full-length murine VEGF-R2 (4085 base pairs (bp)) and C-terminal-truncated VEGF-R2(Δ450C), 2750 bp as follows: 5′ *Not*I primer corresponds to nucleotides 211–232 of VEGF-R2: GCGGCCGC<u>GCGGCCGC</u>AAGGAGTCTGTGCCTGAGAACT. The 3′ *Xho*I primer for full-length VEGF-R2 was CTCGAG<u>CTCGAG</u>AACAGGAGGTGAGCGCAG; the 3′ *Xho*I primer for truncated VEGF-R2 was CTCGAG<u>CTCGAG</u>TAGGAGGTTCACCACATTGAGA. Ligation of purified PCR products was into a *Not*I/*Xho*I site of pShuttle-IRES-hrGFP2 (Stratagene) in-frame with a triple-HA epitope tag. All constructs were subjected to DNA sequencing. Each shuttle vector was linearized with *Pme*I and electroporated into *Escherichia coli* BJ5183 cells for homologous recombination and selection. High-titre adenovirus was generated in HEK293T cells (ATCC) through the transfection of 5 μg, linearized recombinant pAd-1 plasmid with Lipofectamine 2000 (Invitrogen) according to published methods[24].

Site-directed mutagenesis (for tyrosine-to-phenylalanine mutations) required sub-cloning of a 4.1-kilobase *Not*I/*Xho*I fragment containing the VEGF-R2 open reading frame into pBluescript SK- (Stratagene). All results were verified by DNA sequencing, western blotting and flow-cytometry. The following primer pairs (mutant nucleotide exchange is underlined) were used in the mutagenesis reactions using a QuickChange kit (Stratagene) and an annealing temperature of 68 °C: Y1052F sense (GCTTGGCCCGGGACAT<u>TTT</u>TAAAGACCCGGATT) and antisense (CGAACCGGGCCCTGTA<u>AAA</u>TT<u>TC</u>TGGGCCTAA); Y1057F sense (CCGGGACATTTATAAAGACCC<u>GGA</u>T<u>TTT</u>GTCAGAAAAGGAG) and antisense (GGCCCTGTAAATATTTCTGGGC<u>CTA</u><u>AAA</u>CAGTCTTTTCCTC); Y1173F sense (CAGCAGGATGGCAAAGAC<u>TTT</u>A<u>TT</u>GTTCTTCCAATGTCA) and antisense (TGACATTGGAAGAACAAT<u>AAA</u>GTCTTTGCCATCCTGCTG); Y1212F sense (AGTGTGCGACCCCAAAT<u>TCC</u>AT<u>TTT</u>GACAACACAGC) and antisense (GCTGTGTTGTC<u>AAA</u>ATGGAATTTGG<u>GG</u>TCGCACACT).

VEGF receptor expression (semi-quantitative RT–PCR). Total cellular RNA from HUVECs, VSMCs and NHDFs was extracted using Trizol Reagent (Invitrogen). Reverse transcription was performed with a Superscript II Two-Step RT–PCR Kit (Invitrogen). The resultant cDNA template used for PCR reactions included the following primers for human gene products: VEGF-R1 5′-GCGACGTGTGGTCTTACG, 5′-GGCGACTGCAAAAGTCCT; VEGF-R2 5′-CATCCAGTGGGCTGATGA, 5′-TGCCACTTCCAAAAGCAA; VEGF-R3 5′-GATGCGGGACCGTATCTG, 5′-ATCCTCGGAGCCTTCCAC; GAPDH 5′-AAGGGCATCCTGGGCTAC, 5′-GTGGAGGAGTGGGTGTCG. For murine gene products, the following primers were used for PCR reactions: VEGF-R1 5′-GGAGGAAGAGGATGCGGCA, 5′-GAAGCTGTGCTATGTTGCTCTA; VEGF-R2 5′-TTCCAGATGCTGGGCAAGTC, 5′-ATGACATCTTGATTGTGGCAT; VEGF-R3 5′-TGCATGCTGGGTGGACTATCA, 5′-GCAGGAGGAGGAAGAGGAGC; β-actin 5′-GGAGGAAGAGGATGCGGCA, 5′-GAAGCTGTGCTATGTTGCTCTA.

Proximity ligation assay. PLA was performed to image protein–protein interactions using microscopy. In brief, oligonucleotide conjugated 'probe' antibodies are directed against primary antibodies raised against cell-surface receptors. Annealing of the 'probes' occurs when the target proteins are in close proximity, which initiates the amplification of a Texas red reporter signal. VSMCs grown on coverslips were starved for 18 h and treated with 20 ng ml^{-1} PDGF, 40 ng ml^{-1} VEGF or both for 7 min. Cells were washed with chilled PBS and fixed with 4% paraformaldehyde for 15 min, blocked and permeabilized with 5% goat serum (MP Biomedicals) containing 0.2% Triton, and incubated overnight with anti-VEGF-R2 (sc-504) and anti-PDGF-Rβ (sc-432). Proximity ligation was performed according to the manufacturer's protocol[25,26] using the Duolink Detection Kit with PLA PLUS and MINUS Probes for mouse and rabbit (Olink Bioscience). Hoechst stain was included in the Duolink Detection Kit while anti-Phalloidin Alexa488 at 1:5,000 (Invitrogen) was added during the detection reaction. Specimens were mounted with Vectashield mounting media (Vector Laboratories) and examined with an epi-fluorescence microscope (Olympus IX-51) under a ×60 oil objective. Texas-red signal amplification was skeletonized in only the red channel in Metamorph Imaging Software 7.1 (Molecular Devices) for assessment of pixel counts after applying a constant threshold value of −40 (termed 'signal' in the text). Four ×60 oil fields were randomly chosen for analysis and averaged per condition examining four separate samples separately.

Flow cytometry. Fluorescence-activated cell sorting with a FACSCaliber (BD) was performed using standard techniques on live cells using anti-PDGF-Rα (sc-338), anti-PDGF-Rβ (sc-432) and anti-VEGF-R2 (BD) for murine VEGF-R2. Alexa647-conjugated antibodies were used for detection (Invitrogen).

MLEC isolation. MLECs were isolated from 15-week-old C57Bl/6 mice. Briefly, the lungs were excised from euthanized mice, minced and digested with 1 mg ml^{-1} collagenase A/dispase (Roche) in RPMI media. Digested tissue was filtered through a 100-μm nylon cell strainer (BD), and the cell suspension was plated onto collagen-coated dishes. After 3 days, endothelial cells were isolated by immunoselection using Dynal magnetic beads (Invitrogen) conjugated with 5 μg anti-mouse CD-31 and CD-102. Cells were propagated in EBM-2 media supplemented with EBM-2 microvascular SingleQuots (Lonza).

CAM angiogenesis assay. Chick CAM experiments have been previously described[20]. Inhibitors were applied to CAMs 24 h before treatment with growth factor.

Murine Matrigel angiogenesis assay. Murine Matrigel angiogenesis experiments were performed for 7 days in C57/BL6 mice under University of California, San Diego institutional approval using growth-factor-depleted Matrigel (BD) as previously described[21]. Inhibitors or adenovirus were applied with growth factors. Matrigel lysates were generated and homogenized in the RIPA buffer described above. For perfusion where indicated, 100 μL of Griffonia Lectin was injected by tail vein 5 min before Matrigel excision.

Murine fibrosarcoma model. The procedure for generating wild-type and VEGF-null fibrosarcomas has been previously reported[19]. Briefly, VEGF^{+f}/$^{+f}$ murine embryonic fibroblasts (MEFs) were infected with an adenovirus expressing Cre recombinase to delete exon 3 of the VEGF gene. A total of 5×10^6 VEGF-null or wild-type MEFs were injected subcutaneously into C57/Bl6 mice and harvested after 9 days.

Murine orthotopic pancreatic cancer model. The procedure for orthotopic pancreatic tumour implantation was performed as previously described[27]. Briefly, under intraperitoneal ketamine/xylazine anaesthesia, a midline abdominal incision was made and 1×10^6 fast growing (FG) pancreatic cancer cells were implanted by surgical orthotopic implantation into the pancreatic tail of 6- to 8-week-old nude mice with a 1-ml 27G1/2 latex-free syringe (Becton Dickinson). The incision in the abdominal wall was closed with a 6-0 surgical suture in one layer. Fourteen days after tumour implantation, mice were anaesthetized with intraperitoneal ketamine/xylazine and the peritoneal cavity was re-entered through the previous incision. The tumour mass was excised and snap frozen in liquid nitrogen while mice were euthanized by cardiac excision. Tumour lysates were mechanically homogenized in RIPA buffer supplemented with protease inhibitors (Minitab, Roche), 1 mM NaOV, and 10 mM PMSF and 1 mg of protein were used for co-immunoprecipitation as described previously for cultured cells.

Image analysis of vascular histological markers and transgene expression. Frozen tissue sections were used to generate 5-μm cross-sections fixed in 4% paraformaldehyde. Sections were blocked in 5% goat serum (MP Biomedicals) and incubated with primary antibodies overnight. Analysis of pericyte and endothelial cell markers was performed using antibodies to CD-31 (BD), VEGF-R2 (BD), α-SMA (Sigma), PDGF-Rβ (sc-432), p-PDGF-Rβ (Y751, sc-12906-R) and HA-epitope tag (Sigma). AlexaFluor-conjugated secondary antibodies (Invitrogen) were applied and slides were coverslipped with Vectashield mounting media (Vector Labs). Images were acquired using laser scanning confocal microscopy under a ×10 objective (Nikon C1si, Nikon Instruments). CD-31 quantification was performed by acquiring pixel areas from four random fields per condition with Metamorph software. Co-localization of pericytes markers with CD-31$^+$ endothelial cells was performed on similar images thresholded uniformly between experiments at values 40–60 and analysed using the co-localization function of Metamorph Imaging Software 7.1 (Molecular Devices) to identify pixels that had fluorescence intensity in both pericyte marker and endothelial cell marker channels equal to or greater than a set threshold value. The amount of co-localization, as assessed in four randomly selected ×10 fields as measured by overlapping pixel area, was expressed as a percentage. VEGF-R2 transgene expression was similarly assessed by comparing percentage overlap between ectopic truncated VEGF-R2 (by using anti-HA epitope HA tag) and endogenous VEGF-R2 (by using anti-VEGF-R2 raised against the C terminus, sc-504, which does not recognize truncated VEGF-R2Δ450C) in four randomly selected ×10 fields.

Statistics. One-way ANOVA with Bonferroni post-testing compared results between growth-factor treatments. Two-way ANOVA was used for data analysis when genetic or pharmacological interventions were used. An unpaired Student's *t*-test was used in comparisons of VEGF$^{+/+}$ and VEGF$^{-/-}$ tumours.

22. Fong, T. A. T. *et al.* SU-5416 is a potent and selective inhibitor of the vascular endothelial growth factor receptor (Flk-1/KDR) that inhibits kinase catalysis, tumor vascularization, and growth of multiple tumor types. *Cancer Res.* **59**, 99–106 (1999).
23. Klemke, R. L. *et al.* Regulation of cell motility by mitogen-activated protein kinase. *J. Cell Biol.* **137**, 481–492 (1997).
24. Luo, J. *et al.* A protocol for rapid generation of recombinant adenoviruses using the AdEasy system. *Nature Protoc.* **2**, 1536–1547 (2007).
25. Söderberg, O. *et al.* Direct observation of individual endogenous protein complexes *in situ* by proximity ligation. *Nature Methods* **3**, 995–1000 (2006).
26. Jarvius, M. *et al.* In situ detection of phosphorylated platelet-derived growth factor receptor β using a generalized proximity ligation method. *Mol. Cell. Proteomics* **6**, 1500–1509 (2007).
27. Amoh, Y. *et al.* Hair-follicle-derived blood vessels vascularize tumors in skin and are inhibited by doxorubicin. *Cancer Res.* **65**, 5352–5357 (2005).

Vol 456 | 11 December 2008 | doi:10.1038/nature07445

LETTERS

Deletion of vascular endothelial growth factor in myeloid cells accelerates tumorigenesis

Christian Stockmann[1], Andrew Doedens[1], Alexander Weidemann[1], Na Zhang[1], Norihiko Takeda[1], Joshua I. Greenberg[2], David A. Cheresh[3] & Randall S. Johnson[1]

Angiogenesis and the development of a vascular network are required for tumour progression, and they involve the release of angiogenic factors, including vascular endothelial growth factor (VEGF-A), from both malignant and stromal cell types[1]. Infiltration by cells of the myeloid lineage is a hallmark of many tumours, and in many cases the macrophages in these infiltrates express VEGF-A[2]. Here we show that the deletion of inflammatory-cell-derived VEGF-A attenuates the formation of a typical high-density vessel network, thus blocking the angiogenic switch in solid tumours in mice. Vasculature in tumours lacking myeloid-cell-derived VEGF-A was less tortuous, with increased pericyte coverage and decreased vessel length, indicating vascular normalization. In addition, loss of myeloid-derived VEGF-A decreases the phosphorylation of VEGF receptor 2 (VEGFR2) in tumours, even though overall VEGF-A levels in the tumours are unaffected. However, deletion of myeloid-cell VEGF-A resulted in an accelerated tumour progression in multiple subcutaneous isograft models and an autochthonous transgenic model of mammary tumorigenesis, with less overall tumour cell death and decreased tumour hypoxia. Furthermore, loss of myeloid-cell VEGF-A increased the susceptibility of tumours to chemotherapeutic cytotoxicity. This shows that myeloid-derived VEGF-A is essential for the tumorigenic alteration of vasculature and signalling to VEGFR2, and that these changes act to retard, not promote, tumour progression.

To test the role of myeloid-cell-derived VEGF in tumour progression, we created an *in vivo*, cell-lineage-specific, targeted deletion of *Vegfa* by means of crosses of the *loxP*-flanked *Vegfa* allele[3] to the lysozyme M promoter-driven Cre recombinase[4]; this expression is specific to cells of the myeloid lineage, including neutrophils and macrophages, but not dendritic cells (*LysMCre*/*VEGF*+^{f}/+^{f})[4]. This results in *Vegfa* gene excision in about 75% of isolated neutrophils, peritoneal macrophages[5] and tumour-associated macrophages (Supplementary Fig. 1a).

To determine the function of myeloid-cell-derived VEGF-A in an autochthonous mouse model of breast cancer, we then crossed these alleles to the MMTV-*PyMT* transgenic mouse strain[6]. To mitigate the influence of strain variation, individual transgenic alleles were backcrossed to a more than 99% C57Bl/6J strain background (as assayed by single-nucleotide-polymorphism analysis).

Because both genotypes carry the *PyMT* transgene, mice with a myeloid cell-specific deletion of *Vegfa* (MMTV-*PyMT*/*LysMCre*+/*VEGF*+^{f}/+^{f}) will be termed mutant mice, whereas Cre-negative mice (MMTV-*PyMT*/*LysMCre*−/*VEGF*+^{f}/+^{f}) will be termed wild-type (WT) mice (Supplementary Fig. 1b).

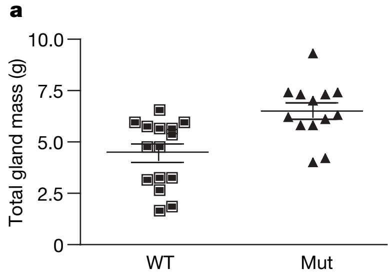
To determine the time point of tumour onset, mammary glands of virgin WT and mutant mice were palpated once a week. The first palpable tumour occurred at an average age of 11 weeks (Supplementary Fig. 1d). Tumours were removed at two different time points, and the total tumour burden for each mouse was evaluated. At 16 weeks, tumours from WT and mutant animals showed no difference in total tumour mass (Supplementary Fig. 1e). However, at 20 weeks, mutant mice had a significantly higher tumour burden than their WT littermates (Fig. 1a). At this point, tumours from WT animals had increased their total mass by almost 400% relative to the 16-week time point; tumours from mutants showed an increase in tumour mass of almost 600% in comparison with the 16-week time point.

A quantitative analysis of proliferating cells showed that the number of cells staining positive for proliferating-cell nuclear antigen (PCNA) increased with progression to malignancy (Supplementary Fig. 1h) and that loss of myeloid-cell-derived VEGF-A increased proliferation rates significantly in the early stages of malignant transformation (Supplementary Fig. 1h). Immunohistological detection of polyoma middle T antigen and PCNA showed that the proliferating cell types were primarily mammary epithelium (Supplementary Fig. 1f).

To address the question of whether the level of malignant progression mirrored tumour mass results, we assessed tumour progression histologically for each time point on the basis of representative sections from each mammary gland by using the criteria described previously[7], in a blinded fashion. As depicted in Fig. 1b, at the time point of 16 weeks, most mammary gland tissues from both WT and mutant animals were still at premalignant stages (96.2% and 89.5%, by area, respectively). After 20 weeks, most tissues in mutant mice had progressed to malignant stages (early carcinoma 38.0%; late carcinoma 52.5%). In WT animals, mammary tissues were primarily at premalignant (30.9%) and early carcinoma (50.2%) stages; a minority of the tissue (18.9%) possessed late-carcinomatous lesions (Fig. 1b). These results indicated that tumours develop more rapidly in the absence of macrophage-derived VEGF-A.

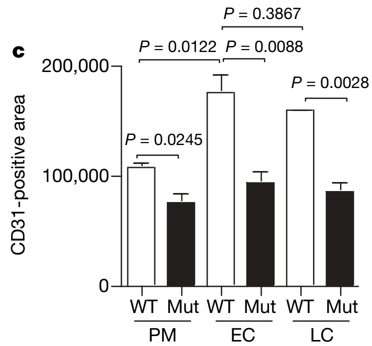
Quantitative analysis of vessel density was then performed (Fig. 1c–e and Supplementary Fig. 1g). In WT mice, premalignant lesions have a low vascular density, which is decreased in mutant mice lacking myeloid VEGF-A (Fig. 1c). Increased vascular density was found in tumours that had progressed to the malignant, early carcinoma stage, which is consistent with the occurrence of an 'angiogenic switch' during the transition to malignancy (Fig. 1c). In late-carcinoma-stage tumours formed by multiple solid nodular areas, the overall vascular density was similar to that seen in the early carcinoma stage in WT animals. However, in mutant mice, the malignancy-associated increased vascularization seen in WT animals did not occur (Fig. 1c). The length and the tortuosity of tumour blood vessels were also affected by the mutation: as depicted in Fig. 1d, the average vessel length in tumours in WT mice increased as tumours progressed to malignancy, indicating that vessel elongation has a key function in the formation of a tumour vessel network in

[1]Molecular Biology Section, Division of Biological Sciences, [2]Department of Surgery, and [3]Department of Pathology and Moores Cancer Center, University of California, San Diego, San Diego, California 92093, USA.



b

Age (wk)	PyMT status	PM	EC	LC
16	WT	96.2±2.4	2.5±2.5	1.3±1.3
	Mut	89.5±7.5	2.2±2.2	8.3±8.3
20	WT	30.9±2.8	50.2±8.8	18.9±6.5
	Mut	9.5±2.8***	38.0±3.6	52.5±4.5**



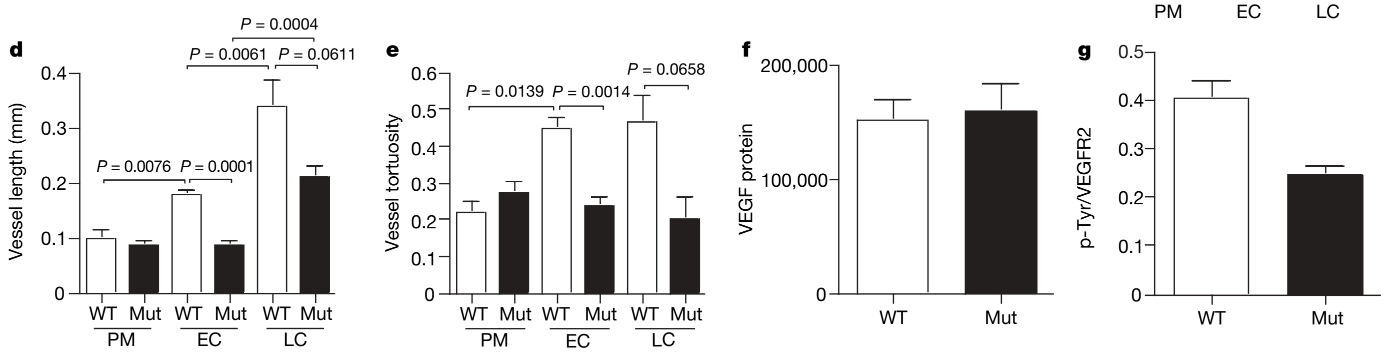


Figure 1 | Deletion of VEGF in myeloid cells results in reduced vascularization but accelerated progression of mammary tumours. a, Total tumour mass of PyMT-WT mice ($n = 15$) and PyMT-mutant mice ($n = 13$) at the age of 20 weeks. $P = 0.0018$.The horizontal lines show mean values. **b**, Distribution of PyMT mammary tumours at prototypical premalignant (PM) lesions, malignant early carcinoma (EC) and late carcinoma (LC) stages in percentages (means ± s.e.m.) between genotypes and at the age of 16 weeks ($n = 4$) and 20 weeks (PyMT-WT $n = 4$, PyMT-mutant $n = 9$). Two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. **c**, Quantitative analysis of the area in pixels covered by CD31-positive cells within a tumour section for each stage ($n = 4$). **d**, Development of vessel length in PyMT-tumours during malignant progression as determined by tracing CD31-positive vessels that were exposed in a longitudinal cut ($n = 4$). **e**, Analysis of vessel tortuosity based on CD31-stained tumour sections ($n = 4$). **f**, Quantitative analysis by photon emission of the VEGF signal (PyMT-WT $n = 7$, PyMT-mutant $n = 4$). $P = 0.7678$. **g**, Ratio of phosphotyrosine (p-Tyr) and VEGFR2 signal intensities as a measure of receptor activation (PyMT-WT $n = 7$, PyMT-mutant $n = 4$). $P = 0.0102$. Error bars show s.e.m.

this model. Along with this increase, blood vessel tortuosity more than doubled in malignant lesions in WT mice (Fig. 1e). In tumours within mutant mice, there was no increase in vessel length before the late carcinoma stage (Fig. 1d), and vessel tortuosity never changed in the mutants (Fig. 1e).

Gene-expression analysis of lysates of mammary tumours from mice at the endpoint age of 20 weeks did not reveal any increase in expression of hypoxia markers, such as *Pgk1*, encoding phosphoglycerate kinase 1 (Supplementary Fig. 1i). However, this analysis of tumours did demonstrate that there was higher expression of *Vegfa* messenger RNA in tumours from mutant animals (Supplementary Fig. 1j). This was not mirrored by changes in VEGF-A protein, however, which did not vary in tumour lysates from WT and mutant animals (Fig. 1f).

We next wished to determine whether signalling to the tumour endothelium was affected by the loss of myeloid-derived VEGF-A. VEGFR2 is a primary endothelial cell-specific receptor tyrosine kinase that participates in VEGF signalling. By immunoprecipitating VEGFR2 from tumour lysates and probing with anti-phosphotyrosine followed by anti-VEGFR2 antibody by means of western blotting, we were able to quantify activated and total VEGFR2 from WT and mutant tumours. As shown in Fig. 1g, loss of myeloid-derived VEGF-A causes a roughly 50% decrease in the ratio of phosphorylated VEGFR2 to total VEGFR2 in comparison with WT tumour lysates. Decreased VEGFR2 phosphorylation suggests that myeloid-cell-derived VEGF-A has a unique function in tumour vascularization that cannot be compensated for by VEGF-A from other sources within the tumour.

To analyse these effects in fully transformed cells, we next performed experiments with subcutaneous isografts of Lewis lung carcinoma (LLC) cells. LLC tumours in mutant mice lacking myeloid VEGF had significantly greater tumour volumes at day 12 (Fig. 2a), with no genotype-specific differences in tumour infiltration (Supplementary Fig. 2). Flow cytometric analysis demonstrated that macrophages were the predominant myeloid infiltrate and that levels of infiltrating macrophages and less abundant neutrophils were similar across genotypes (Supplementary Fig. 2).

A significant decrease in vascular density and vessel tortuosity was again observed in mutant mice (Fig. 2b, c). An analysis of VEGF-A protein and *Vegfa* mRNA (Supplementary Fig. 3a, b) showed no significant differences between genotypes. However, when VEGFR2, phosphorylated VEGFR2 and the ratio of phosphorylated VEGFR2 to total VEGFR2 were quantified (Fig. 2d), activation of VEGFR2 was once again decreased by about 50% in mutant tumours relative to WT tumours.

We next wished to determine the relationship between the expression of myeloid cell VEGF and vessel normalization; to assay this, we immunohistologically detected pericyte coverage of tumour vessels and assayed the levels of pericyte coverage in the tumours (Fig. 2e). As shown, the loss of VEGF-A expression in the myeloid cells resulted in a marked increase in the level of vessel pericyte association, indicating that VEGF-A expression from infiltrating myeloid cells is essential for intratumoral loss of vessel pericyte coverage. Vessel permeability also decreased in tumours from mutant animals (Fig. 2f).

Hypoxia is associated with the growth of solid tumours and with tumorigenic vasculature. We found that the more slowly growing WT tumours had higher levels of hypoxia (Fig. 3a and Supplementary Fig. 3c) and that this occurred even in areas that were extensively vascularized (Supplementary Fig. 3c). Staining of tumour sections for VEGF and F4/80, a macrophage marker, demonstrates that intense VEGF-A immunostaining in WT tumours occurs in regions of high infiltration, whereas in mutant tumours VEGF-A staining is only found outside of infiltrated regions (Supplementary Fig. 3d); this shows how VEGF-A from myeloid cells could have an acute and localized effect on tumour vasculature.

To determine whether the vascular changes we describe confer an increased susceptibility to cytotoxic agents, we administered the chemotherapy agent cyclophosphamide to WT and mutant LLC-

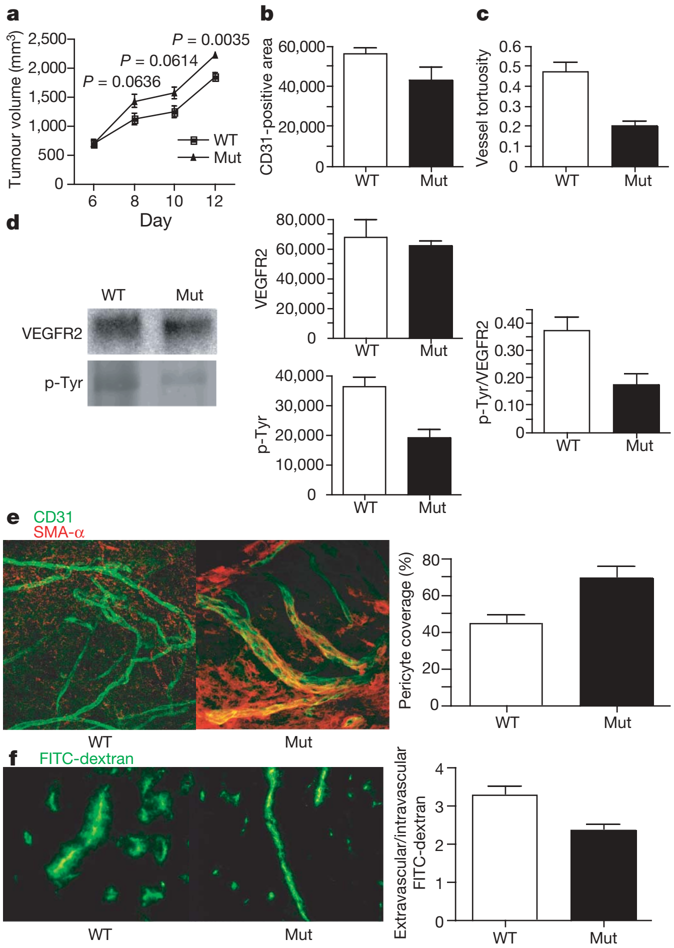


Figure 2 | Deletion of VEGF in myeloid cells leads to a normalized vasculature and higher tumour volumes in LLC. **a**, Growth-curve analysis of LLC tumours injected subcutaneously into WT and mutant mice ($n > 7$ for each group). **b**, Quantitative analysis of pixels corresponding to CD31-positive endothelial cells ($n = 4$). $P = 0.0406$. **c**, Determination of blood vessel tortuosity in LLC tumours. $P = 0.0022$. **d**, Left: immunoblotting for VEGFR2 and phosphotyrosine (p-Tyr) after immunoprecipitation of VEGFR2 from LLC tumour lysates. Centre: quantitative analysis of VEGFR2 (upper; $P = 0.7460$) and phosphotyrosine (lower; $P = 0.0304$) signals as measured by photon emission. Right: ratio of p-Tyr to VEGFR2 signal intensities as a measure of receptor activation (WT $n = 8$, mutant $n = 7$); $P = 0.0075$. **e**, Left: confocal microscopy images of simultaneous immunodetection of endothelial cells and pericytes in LLC tumours with the specific markers CD31 and smooth muscle actin-α (SMA-α). Right: pericyte coverage as assessed by SMA-α/CD31 co-localization ($n = 4$). $P = 0.0201$. **f**, Left: fluorescent microscopy images of fluorescein isothiocyanate (FITC)-dextran angiography on LLC isografts. Right: ratio of extravascular to intravascular FITC-dextran as a measure of vascular permeability (WT $n = 6$, mutant $n = 4$). $P = 0.0217$. Error bars show s.e.m.

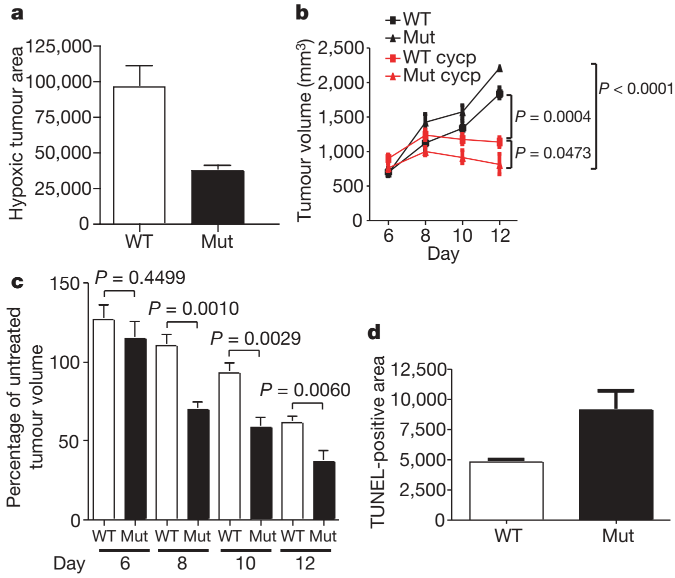


Figure 3 | Deletion of VEGF in myeloid cells results in reduced hypoxia and increased susceptibility of LLC tumours to cytotoxic agents. **a**, Quantitative analysis of pixels corresponding to hypoxic tumour areas (WT $n = 6$, mutant $n = 4$). $P = 0.0112$. **b**, Growth-curve analysis of LLC isografts from WT and mutant animals ($n > 4$ for each group) treated with cyclophosphamide (cycp; 170 mg kg^{-1}) at days 6, 8 and 10 after tumour implantation. **c**, Response of tumours from WT and mutant mice to treatment with cyclophosphamide, expressed as the percentage of treated tumour volume to untreated tumour volume at the specified time points. **d**, Quantitative analysis of TdT-mediated dUTP nick end labelling (TUNEL) positive area on cyclophosphamide-treated LLC isografts ($n = 4$). $P = 0.0343$. Error bars show s.e.m.

tumour-bearing mice. Loss of myeloid-derived VEGF increased the chemotherapeutic efficacy of the treatment (Fig. 3b, c). The cyclophosphamide treatment also resulted in a significantly increased level of tumour cell death in mice bearing tumours lacking myeloid VEGF-A, as can be seen in Supplementary Fig. 3e and as quantified in Fig. 3d. Similar experiments performed with the cytotoxic drug *cis*-platinum gave comparable results (Supplementary Fig. 3f, g).

To determine the differential contributions of tumour-cell and myeloid-cell-derived VEGF, we next created fibrosarcoma cell lines derived from C57Bl/6J embryos homozygous for the conditional allele of the *Vegfa* gene. We then used Cre recombinase expression from adenovirus to create matching WT (infected with control adenovirus) and *Vegfa*-null (infected with Cre recombinase-expressing adenovirus) tumour cell lines, and injected them subcutaneously into mice that were WT or null for myeloid VEGF-A expression. As can be seen in Fig. 4a, fibrosarcomas that were WT for VEGF-A expression grew faster and significantly larger in mice lacking *Vegfa* in myeloid cells, which is similar to what we observed in the other models described above. However, in agreement with previous findings from our laboratory[8], tumours lacking VEGF-A expression grew more slowly than those expressing VEGF-A, when implanted into WT mice (Fig. 4a).

When VEGF-A was absent from both the tumour and the myeloid compartment, tumour growth essentially collapsed, with greatly increased apoptosis and necrosis (Fig. 4a–f). Further, although there was a decrease in tumour vessel density (Fig. 4b, d) ranging from the highest vessel count in tumours that were WT in all tissue compartments, to the lowest in those lacking VEGF-A in tumour cells and myeloid cells, these values did not correspond directly to changes in tumour growth; rather, they demonstrate the unique role for myeloid-derived VEGF-A in facilitating changes in tumour vessel function and normalization.

Tumour-associated macrophages have been shown to be crucial for malignant progression in a mouse model of breast cancer[7,9]; however, these studies removed macrophages completely from tumours and were therefore removing all of the many aspects of macrophage involvement in tumour progression. In isolating one of those aspects experimentally, we have shown that myeloid-derived VEGF-A is required for the formation of a high-density vessel network, that loss of VEGF-A from this source greatly decreases the activation of VEGFR2 in spite of little or no loss in total tumour VEGF-A, and that the absence of this increased vascular density does not hamper, but rather accelerates, tumour progression.

When administered as single agents, anti-angiogenic drugs have produced modest objective responses in clinical trials, but overall these single-agent approaches have not yielded significant long-term survival benefits[10]. In contrast, when given in combination with chemotherapy, antibodies targeted against VEGF-A result in an increase in survival in colorectal cancer patients[11]. Our observations

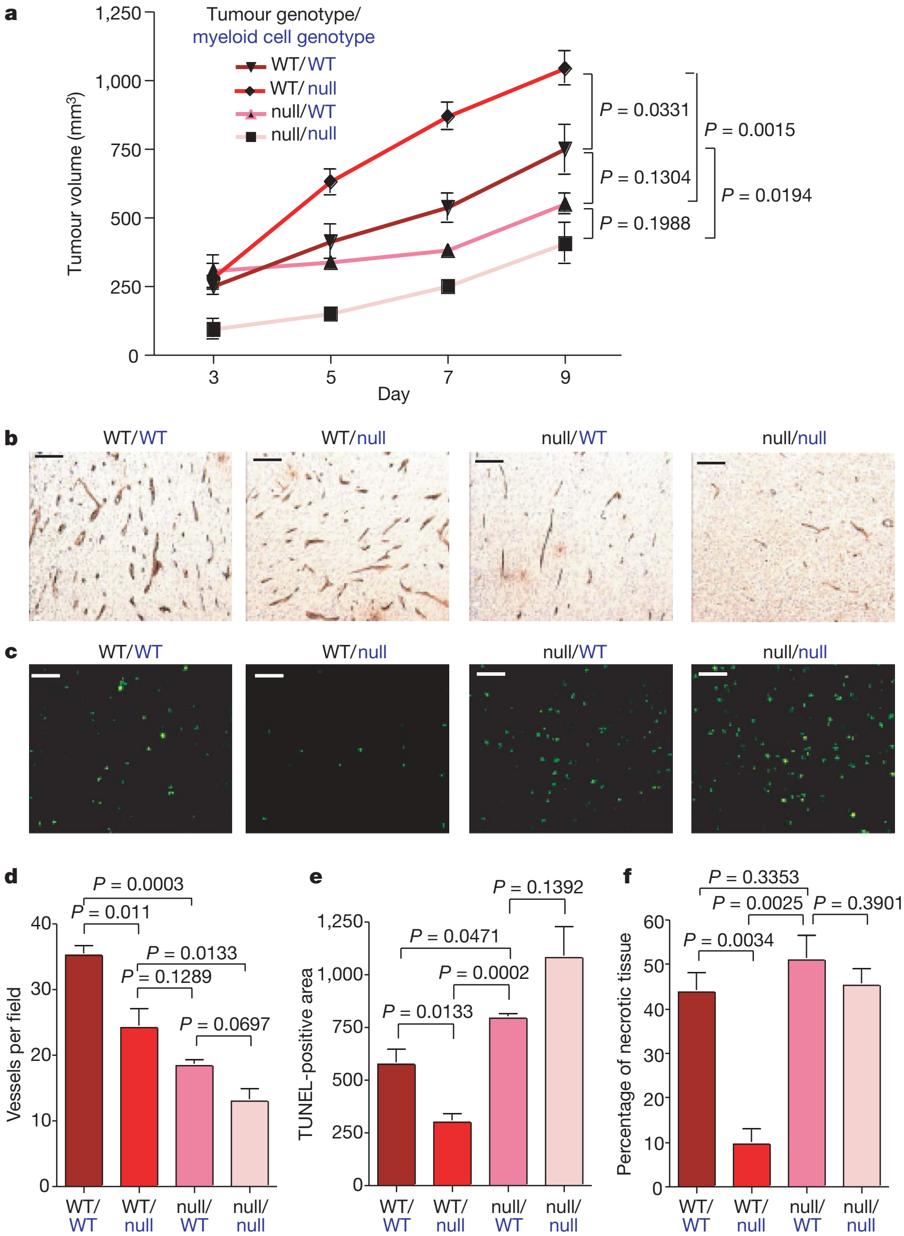


Figure 4 | Effect of tumour-cell-derived versus myeloid-cell-derived VEGF on tumour angiogenesis and growth. a, Growth-curve analysis of wild-type (WT) and VEGF nullizygous (null) fibrosarcoma isografts (genotype labelled in black) implanted into WT mice or mutant mice (null) with a myeloid-cell-specific deletion of *VEGF* (genotype labelled in blue) ($n = 4$ for each group). **b**, CD34 immunostaining on fibrosarcoma isografts. **c**, Detection of apoptotic cells in fibrosarcomas by TUNEL staining. **d**, Quantitative analysis of CD34-positive blood vessels ($n = 4$). **e**, Quantification of TUNEL-positive cells ($n = 4$). **f**, Assessment of tumour necrosis on fibrosarcoma midline sections. Shown is the perimeter of necrotic areas expressed as a percentage of the total tumour perimeter ($n = 4$). Scale bars, 100 μm. Error bars show s.e.m.

argue that a critical mediator of this sensitization to chemotherapy lies in the infiltrating innate immune cells rather than in the malignant cells of the tumour. These findings open a new avenue for the investigation of anti-angiogenic therapies targeted to tumour-associated cells of the immune system.

METHODS SUMMARY

Transgenic mice (C57Bl/6J) expressing the polyoma middle T (PyMT) oncoprotein under the promoter of the mouse mammary tumour virus (MMTV) long terminal repeat were bred to mice (C57Bl/6J) with both alleles of exon 3 of *Vegfa* flanked by *loxP* sites (*VEGF*+^{f}/+^{f})[12]. Myeloid cell-specific knockout of VEGF was achieved by breeding male mice homozygous for the floxed *Vegfa* allele and hemizygous for the PyMT oncogene (MMTV-*PyMT*/*VEGF*+^{f}/+^{f}) with female mice (C57Bl/6J) homozygous for the floxed *Vegfa* allele expressing Cre recombinase driven by the lysozyme M promoter (*LysMCre*+/*VEGF*+^{f}/+^{f})[4]. For our studies, we used female mice hemizygous for the *PyMT* oncogene carrying two floxed *Vegfa* alleles and positive for Cre expression (MMTV-*PyMT*/*LysMCre*+/*VEGF*+^{f}/+^{f}), designated as mutants, whereas female littermates negative for Cre expression (MMTV-*PyMT*/*LysMCre*−/*VEGF*+^{f}/+^{f}) served as WT controls. Loss of VEGF-A expression in myeloid cells had no effect on normal mammary development (Supplementary Fig. 1c).

To generate isografts, 10^7 LLC cells on a BL6 background (ATCC) were injected subcutaneously into *LysMCre*−/*VEGF*+^{f}/+^{f} (WT) and *LysMCre*+/*VEGF*+^{f}/+^{f} mice (mutant). Mouse embryonic fibroblasts (MEFs) were isolated from WT mice. The transgenic MEFs were immortalized by stable transfection with SV40 large T antigen and then transformed with a vector expressing oncogenic mutant H-Ras (Val 12). Subsequently, the *VEGF*+^{f}/+^{f} MEFs were infected with an adenovirus expressing Cre recombinase to delete exon 3 of the *Vegfa* gene. A total of 5×10^6 VEGF-A-null or WT MEFs were injected subcutaneously into WT and mutant mice. Data are expressed as means ± s.e.m. Statistical significance was determined by a two-tailed unpaired *t*-test.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 March; accepted 12 September 2008.
Published online 9 November 2008.

1. Folkman, J. Angiogenesis. *Annu. Rev. Med.* **57,** 1–18 (2006).
2. Lewis, J. S., Landers, R. J., Underwood, J. C., Harris, A. L. & Lewis, C. E. Expression of vascular endothelial growth factor by macrophages is up-regulated in poorly vascularized areas of breast carcinomas. *J. Pathol.* **192,** 150–158 (2000).
3. Gerber, H. *et al.* VEGF couples hypertrophic cartilage remodeling, ossification and angiogenesis during endochondral bone formation. *Nature Med.* **5,** 623–628 (1999).
4. Clausen, B. E., Burkhardt, C., Reith, W., Renkawitz, R. & Forster, I. Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res.* **8,** 265–277 (1999).
5. Cramer, T. *et al.* HIF-1α is essential for myeloid cell-mediated inflammation. *Cell* **112,** 645–657 (2003).
6. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol. Cell. Biol.* **12,** 954–961 (1992).
7. Lin, E. Y. *et al.* Progression to malignancy in the polyoma middle T oncoprotein mouse breast cancer model provides a reliable model for human diseases. *Am. J. Pathol.* **163,** 2113–2126 (2003).
8. Grunstein, J., Roberts, W. G., Mathieu-Costello, O., Hanahan, D. & Johnson, R. S. Tumor-derived expression of vascular endothelial growth factor is a critical factor in tumor expansion and vascular function. *Cancer Res.* **59,** 1592–1598 (1999).
9. Lin, E. Y. *et al.* Macrophages regulate the angiogenic switch in a mouse model of breast cancer. *Cancer Res.* **66,** 11238–11246 (2006).
10. Mayer, E. L., Lin, N. U. & Burstein, H. J. Novel approaches to advanced breast cancer: bevacizumab and lapatinib. *J. Natl. Compr. Canc. Netw.* **5,** 314–323 (2007).
11. Hurwitz, H. I., Honeycutt, W., Haley, S. & Favaro, J. Long-term treatment with bevacizumab for patients with metastatic colorectal cancer: case report. *Clin. Colorectal Cancer* **6,** 66–69 (2006).
12. Gerber, H. P. *et al.* VEGF is required for growth and survival in neonatal mice. *Development* **126,** 1149–1159 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. DuRose, N. Ferrara, L. Iruela-Arispe, L. Coussens, M. Karin, V. Nizet and H.-P. Gerber for helpful conversations. Financial support was provided by the Deutsche Forschungsgemeinschaft to C.S. (STO 787/1-1) and A.W. (WE 4275/1-1), a Susan G. Komen Dissertation Research Award (DISS0402406) to A.D., and National Institutes of Health grants CA82515, CA118165 and AI060840 to R.S.J.

Author Contributions C.S. generated isografts, conducted drug treatment studies, performed biochemical experiments, immunohistochemical procedures and histological analysis, and prepared the manuscript. A.D. generated MMTV-*PyMT/LysMCre/VEGF*$+^{f}/+^{f}$ mice, conducted related tumour palpation/mass studies, initial histology and flow cytometry, and assisted with the manuscript. A.W. generated VEGF-null fibroblasts. N.Z. performed preliminary cDNA studies on MMTV-*PyMT/LysMCre/VEGF*$+^{f}/+^{f}$ tumours. N.T. assisted with the fluorescein isothiocyanate-dextran angiography. J.I.G. and D.A.C. provided confocal microscopy images and technical advice. R.S.J. supervised and directed the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.S.J. (rsjohnson@ucsd.edu).

METHODS

Chemotherapy. The treatment with chemotherapy was started six days after injection of 10^7 LLC cells. Cisplatin ($5\,mg\,kg^{-1}$) and cyclophosphamide ($170\,mg\,kg^{-1}$) were purchased from Sigma and given intraperitoneally on alternate days until day 10.

Histology, immunohistochemistry and immunofluorescence. For whole-mount preparation of mammary glands, the inguinal gland was removed and fixed in 10% phosphate-buffered formalin (Fisher Scientific) and dehydrated through a series of graded acetone and ethanol washes. Glands were stained for 3 h with Harris haematoxylin (VWR International), and excess stain was removed by washing in water. Glands were destained in acidic 50% ethanol and dehydrated and stored in methyl salicylate (Fisher Scientific).

After removal, tumours were fixed in 10% formalin, zinc fixative (BD Pharmingen) or 4% paraformaldehyde, respectively, and then embedded in paraffin. For haematoxylin/eosin staining and immunostaining, 5-μm sections were deparaffinized with xylene and rehydrated with graded ethanol. The sections were stained in accordance with routine immunohistochemistry procedures and revealed with Vectastain ABC or ABC-AP kit (Vector Laboratories).

Primary antibodies used for immmunohistochemistry and immunofluorescence. Biotinylated rat anti-CD31 at 1:100 dilution (BD Pharmingen), rat anti-CD34 (Novus), biotinylated mouse anti-PCNA at 1:800 dilution (BD Pharmingen), biotinylated rat anti-F4/80 at 1:200 dilution (Serotec), rabbit anti-VEGF at 1:50 dilution (sc-152; Santa Cruz) and mouse anti-SMA-α at 1:500 dilution (Chemicon) were used. Fluorochrome-conjugated Alexa 488 and Alexa 555 were used as secondary antibodies (1:200 dilution).

Pimonidazole hydrochloride (Hypoxyprobe-1) was injected intraperitoneally ($60\,mg\,kg^{-1}$ body weight) 1 h before tumour removal and was detected with a Mouse IgG_1 monoclonal antibody (Mab1; Natural Pharmacia International).

TUNEL assay (Promega) was performed in accordance with the manufacturer's instructions.

Quantitative analysis of histological markers. For quantitative analysis of the distribution of immunohistochemical markers within the tumour, the midline sections of tumours were photographed into TIFF images with a Leica DMR microscope and SPOT RT colour camera system (Diagnostic Instruments Inc.) and the area (number of pixels) with positive staining equal to or greater than a set threshold was measured with the Image J program; such marked areas were normalized by the number of images for each tumour. For comparison at different stages of progression, PyMT tumours were classified as described previously[7].

To determine vessel density, the vasculature marked by CD31 and CD34 was skeletonized with the Image J program; the area covered by blood vessels was measured and the number of vessels per field was determined, respectively.

To estimate change in vessel length, the skeletonized blood vessels that showed up in a longitudinal cut within an image were traced along their midline and the number of pixels was converted into distance in millimetres with Image J.

The tortuosity (T) of blood vessel is defined as $T = L/S - 1$, where L is the length of the vessel of interest and S is the straight-line distance between its endpoints.

Confocal microscopy. Co-localization of pericytes (SMA-α) endothelial cells (CD31) was imaged on 20-μm sections of excised LLC tumours. Images were acquired with laser-scanning confocal microscopy under a 60×/1.4 numerical aperture oil-immersion objective (Nikon C1si; Nikon Instruments Inc.).

Flow cytometry. Tumours were excised and cut into small pieces before incubation in a shaker at 37 °C at 250 r.p.m. for 2 h in 0.1% collagenase A (Roche) in RPMI 1640 medium (Invitrogen). The resulting single-cell suspension was pelleted and rinsed in PBS before treatment with hypotonic lysis buffer and passing through a 40-μm cell strainer (BD Biosciences). For each condition, 10^6 cells were then incubated with Fc Block (BD Biosciences) to block non-specific FcR binding before labelling with fluorescently conjugated antibodies (eBiosciences) with standard protocols. Data were acquired on a FACSCalibur flow cytometer (BD Biosciences) with propidium iodide to exclude non-viable cells, and analysed with FlowJo software (Tree Star).

Determination of vascular permeability. FITC-dextran (200 μl; 2 MDa; Sigma) were injected intravenously at a concentration of $50\,mg\,ml^{-1}$. After 10 min, tumours were removed and immediately frozen in OCT compound. Cryosections 10 μm thick were cut, and vascular permeability was determined on the basis of ten randomly taken pictures from each tumour. With the use of Image J the edges of blood vessels were indentified and the extravascular and intravascular areas (numbers of pixels) covered by FITC-dextran were measured. To determine vascular permeability in a quantitative manner, the ratio of extravascular staining to intravascular FITC-dextran was calculated.

Immunoprecipitation and immunoblotting. Tumours were lysed in RIPA buffer and 500 μg of lysate was used for immunoprecipitation of VEGFR2. The following antibodies were used in this study: rabbit anti-VEGFR2 (55B11; Cell Signaling), horseradish peroxidase-conjugated anti-phosphotyrosine (4G10; Millipore), rabbit anti-VEGF (sc-152; Santa Cruz) and mouse anti-β-actin (Sigma). For quantitative analysis the membranes were scanned with a fluorescence scanner and the signal strength was determined with ImageQuant software.

Determination of deletion frequency and gene expression by quantitative PCR. Thioglycollate-elicited macrophages as well as FACS-sorted F4/80-positive tumour-associated macrophages isolated from PyMT-mutant tumours were used for isolation of genomic DNA as described previously[5]. Whole tumours were ground to a fine powder in liquid nitrogen and subsequently homogenized in TRIzol reagent (Invitrogen). Total RNA was isolated in accordance with the manufacturer's instructions. For real-time PCR analysis, the isolated RNA was reverse transcribed. For PCR reactions, TaqMan-Universal Mastermix (Applied Biosytems) was used. cDNA (50 ng) was used as template to determine the relative amount of mRNA by real-time PCR (ABI Prism 7700 sequence detection system), using specific primers and probes with the following sequences: VEGF, 5′-ATCCGCATGATCTGCATGG-3′ (reverse), and 5′-AGTCCCATGAAGTGATCAAGTTCA-3′ (forward); VEGF probe, (6FAM)TGCCCACGTCAGAGAGCAACATCAC(BHQ-1); PGK, 5′-TTCTTGCTGCTCTCAGTACCACA-3′ (reverse) and 5′-CAAATTTGATGAGAATGCCAAGACT-3′ (forward); PGK probe, (6FAM)TATACCTGCTGGCTGGATGGGCTTGGACT(BHQ-1) (6FAM is 6-carboxyfluorescein; BHQ-1 is Black hole quencher (ABI)).

Cell culture. Cells were cultured in DMEM High Glucose medium, supplemented with 10% fetal bovine serum, 2 mM glutamine, $50\,U\,ml^{-1}$ penicillin and $100\,\mu g\,ml^{-1}$ streptomycin.

LETTERS

DNA double-strand breaks activate a multi-functional genetic program in developing lymphocytes

Andrea L. Bredemeyer[1]*, Beth A. Helmink[1]*, Cynthia L. Innes[2], Boris Calderon[1], Lisa M. McGinnis[1], Grace K. Mahowald[1], Eric J. Gapud[1], Laura M. Walker[1], Jennifer B. Collins[2], Brian K. Weaver[1], Laura Mandik-Nayak[1], Robert D. Schreiber[1], Paul M. Allen[1], Michael J. May[3], Richard S. Paules[2], Craig H. Bassing[4,5] & Barry P. Sleckman[1]

DNA double-strand breaks are generated by genotoxic agents and by cellular endonucleases as intermediates of several important physiological processes. The cellular response to genotoxic DNA breaks includes the activation of transcriptional programs known primarily to regulate cell-cycle checkpoints and cell survival[1–5]. DNA double-strand breaks are generated in all developing lymphocytes during the assembly of antigen receptor genes, a process that is essential for normal lymphocyte development. Here we show that in murine lymphocytes these physiological DNA breaks activate a broad transcriptional program. This program transcends the canonical DNA double-strand break response and includes many genes that regulate diverse cellular processes important for lymphocyte development. Moreover, the expression of several of these genes is regulated similarly in response to genotoxic DNA damage. Thus, physiological DNA double-strand breaks provide cues that can regulate cell-type-specific processes not directly involved in maintaining the integrity of the genome, and genotoxic DNA breaks could disrupt normal cellular functions by corrupting these processes.

The generation of complete antigen receptor genes in developing lymphocytes requires assembly of the second exon through the process of V(D)J recombination[6]. This process is initiated by the recombinase activating gene (Rag) endonuclease, composed of Rag1 and Rag2, which introduces DNA double-strand breaks (DSBs) at the border of two gene segments and their flanking Rag recognition sequences (recombination signals)[7]. Rag cleavage occurs in cells at the G1 phase of the cell cycle, and the resulting pairs of coding and signal DNA ends are processed and joined by the non-homologous end-joining (NHEJ) pathway of DNA DSB repair, generating a coding joint and a signal joint, respectively[8]. All developing lymphocytes must generate and repair several Rag DSBs as they assemble the genes encoding the heterodimeric B and T cell receptors.

The cellular response to genotoxic DSBs is initiated by the ataxia telangiectasia mutated (Atm) serine threonine kinase[1–3]. A central feature of this response is the activation of transcriptional programs, such as those initiated by p53, which primarily function to promote cell-cycle arrest and apoptosis of cells with extensive, or persistent, DNA breaks[9]. Whether physiological DSBs generally activate DSB responses that include these transcriptional programs has been unclear. In this regard, stabilization of p53 was observed only in NHEJ-deficient, and not wild-type, thymocytes undergoing V(D)J recombination, suggesting that Rag DSBs may only activate transcriptional programs in a small fraction of developing lymphocytes in which these breaks are not efficiently repaired and thus pose a hazard to the cell[10]. Indeed, the unopposed activation of p53 in response to transient Rag DSBs could lead to the unwanted apoptosis of developing lymphocytes that would have otherwise undergone successful antigen receptor gene assembly. Alternatively, all Rag DSBs could generally activate transcriptional pathways, including those that promote survival. In this regard, genotoxic DSBs have been shown to activate NF-κB, and, although the genes that are regulated by NF-κB in this setting have not been identified, in other settings NF-κB regulates the expression of genes with anti-apoptotic activity[5,11,12]. Moreover, NF-κB promotes the expression of genes involved in diverse cell-type-specific processes, raising the intriguing possibility that genotoxic, and perhaps physiological, DSBs activate genetic programs with broader cellular functions[13].

Initially we wished to determine whether Rag DSBs are capable of activating NF-κB. To this end, we generated several NHEJ-deficient (*Artemis*$^{-/-}$ (*Artemis* is also known as *Dclre1c*), *Ku70*$^{-/-}$ (*Ku70* is also known as *Xrcc6*) and *Scid* (deficient in *Prkdc*)) v-abl-transformed pre-B cell lines, hereafter referred to as abl pre-B cells. G1 cell-cycle arrest and Rag expression can be induced in these cells using the v-abl kinase inhibitor, STI571 (refs 14 and 15). As these cells are NHEJ-deficient, Rag induction results in the accumulation of unrepaired DSBs at the immunoglobulin (Ig) light (L) chain κ locus (Supplementary Fig. 1 and data not shown). These DSBs lead to activation of NF-κB as evidenced by the increased nuclear translocation of NF-κB (primarily p50/p65) in STI571-treated NHEJ-deficient (*Artemis*$^{-/-}$, *Ku70*$^{-/-}$ and *Scid*) abl pre-B cells as compared to STI571-treated *Rag2*$^{-/-}$ abl pre-B cells (Fig. 1a and Supplementary Fig. 2a and b). This NF-κB activation was partially dependent on Atm as it was significantly diminished in *Artemis*$^{-/-}$ *Atm*$^{-/-}$ abl pre-B cells and in *Artemis*$^{-/-}$ abl pre-B cells that were treated with an Atm inhibitor (Fig. 1a and Supplementary Fig. 2c). As was observed for genotoxic DSBs, NEMO (NF-κB essential modulator) is required for the optimal activation of NF-κB in response to Rag DSBs (Supplementary Fig. 2d)[12,16,17].

To determine whether the activation of NF-κB, and possibly other transcription factors, by Rag DSBs leads to alterations in gene expression, we carried out gene expression profiling on three independently derived *Rag2*$^{-/-}$ and *Artemis*$^{-/-}$ abl pre-B cells treated with STI571 for 48 h (Supplementary Fig. 3). These analyses showed that the expression of 364 genes was changed ≥twofold in all three *Artemis*$^{-/-}$ abl pre-B cells relative to the mean expression level in the three *Rag2*$^{-/-}$ cells (Fig. 1b, Supplementary Fig. 3a and Supplementary Tables 1, 2 and 3). Several of the gene expression changes were confirmed through quantitative real-time PCR (qRT–PCR) and protein expression analyses (Fig. 1c and Supplementary

[1]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [2]Environmental Stress and Cancer Group, and NIEHS Microarray Group, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA. [3]School of Veterinary Medicine, and [4]Department of Pathology and Laboratory Medicine, Center for Childhood Cancer Research, Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia 19104, USA. [5]Abramson Family Cancer Research Institute, Philadelphia, Philadelphia 19104, USA.
*These authors contributed equally to this work.

Figs 4 and 5). These changes were not due to Artemis deficiency *per se*, as they were also observed in response to Rag DSBs generated in *Ku70*$^{-/-}$ and *Scid* (Prkdc-deficient) abl pre-B cells, but not in STI571-treated *Artemis*$^{-/-}$ *Rag2*$^{-/-}$ abl pre-B cells (Supplementary Fig. 6). Thus, Rag DSBs regulate the expression of a large cohort of genes. A significant portion of this cohort is dependent on the activation of Atm by Rag DSBs as evidenced by the analyses of *Artemis*$^{-/-}$ *Atm*$^{-/-}$ abl pre-B cells (Fig. 1, Supplementary Figs 3–5 and 7, and Supplementary Tables 2 and 3). Notably, however, approximately half of the genes are Atm-independent, demonstrating that proteins other than Atm can initiate responses to Rag DSBs in G1-phase cells. Furthermore, analyses of *Artemis*$^{-/-}$ abl pre-B cells expressing a dominant negative amino-truncated form of IκBα (IκBα-ΔN), which inhibits NF-κB activation, showed that 75 of the Atm-dependent gene expression changes were also NF-κB-dependent (Fig. 1b and c, Supplementary Figs 3 and 4, and Supplementary Table 3). As only some of the Rag DSB-dependent genes are dependent on NF-κB, additional transcriptional pathways must be integrated into the Rag DSB response. Thus, Rag DSBs promote a broad genetic program, in part, through the activation of Atm and NF-κB.

A similar genetic program is activated by Rag DSBs in developing B cells *in vivo*. This is evidenced by gene expression profiling of purified B220^{+} IgM^{-} bone marrow B cells, which are primarily pre-B cells, from *Artemis*$^{-/-}$ and *Rag1*$^{-/-}$ mice expressing an IgH transgene (*Artemis*$^{-/-}$ *IgHtg* and *Rag1*$^{-/-}$ *IgHtg*, respectively; Supplementary Fig. 8 and Supplementary Table 4). Of the 364 Rag-DSB-responsive genes identified by the analysis of *Artemis*$^{-/-}$ abl pre-B cells, 109 showed significant changes in expression in *Artemis*$^{-/-}$ *IgHtg* pre-B cells, as compared to *Rag1*$^{-/-}$ *IgHtg* pre-B cells (Fig. 1b, Supplementary Fig. 8 and Supplementary Tables 1–4). The asynchronous nature of Rag DSB induction in developing pre-B cells is probably responsible, in part, for the reduced magnitude and number of the gene expression changes in primary *Artemis*$^{-/-}$ *IgHtg* pre-B cells, as compared to the *Artemis*$^{-/-}$ abl pre-B cells. Markedly, these gene expression changes do lead to significant increases in the expression of the proteins they encode (Fig. 2a and Supplementary Fig. 9). That these genes are Rag DSB-responsive was further demonstrated by the analysis of *Artemis*$^{-/-}$ *IgHtg* and *Rag1*$^{-/-}$ *IgHtg* bone marrow interleukin (IL)-7 pre-B cell cultures. Withdrawal of IL-7 from *Artemis*$^{-/-}$ *IgHtg*, but not from *Rag1*$^{-/-}$ *IgHtg*, bone marrow IL-7 cultures led to an increased accumulation of unrepaired IgLκ coding ends and a robust upregulation in the expression of many of

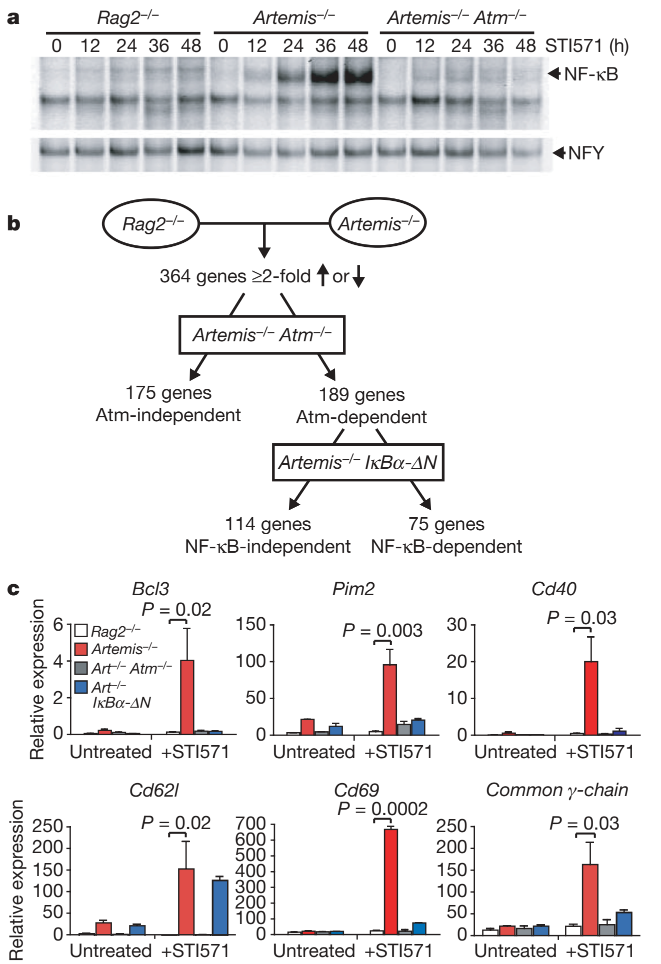


Figure 1 | Rag DSBs activate a broad genetic program. a, NF-κB EMSA of nuclear lysates from *Rag2*$^{-/-}$, *Artemis*$^{-/-}$ and *Artemis*$^{-/-}$ *Atm*$^{-/-}$ abl pre-B cells treated with STI571 for the indicated number of hours. NFY EMSA is shown as a control. Results are representative of three experiments. **b,** Schematic of gene expression changes in response to Rag DSBs by microarray analyses. The Atm-independent gene expression changes are reported in Supplementary Table 1. The Atm-dependent gene expression changes are reported in Supplementary Table 2 (NF-κB-independent) and Supplementary Table 3 (NF-κB-dependent). **c,** qRT–PCR analysis of messenger RNA isolated from *Rag2*$^{-/-}$ (white bars), *Artemis*$^{-/-}$ (red bars), *Artemis*$^{-/-}$ *Atm*$^{-/-}$ (*Art*$^{-/-}$ *Atm*$^{-/-}$, grey bars) and *Artemis*$^{-/-}$ *IκBα-ΔN* (*Art*$^{-/-}$ *IκBα-ΔN*; blue bars) abl pre-B cells treated with STI571 for 0 or 48 h. Mean and s.d. from two experiments are shown; *P* values were calculated using a Student's one-tailed *t*-test.

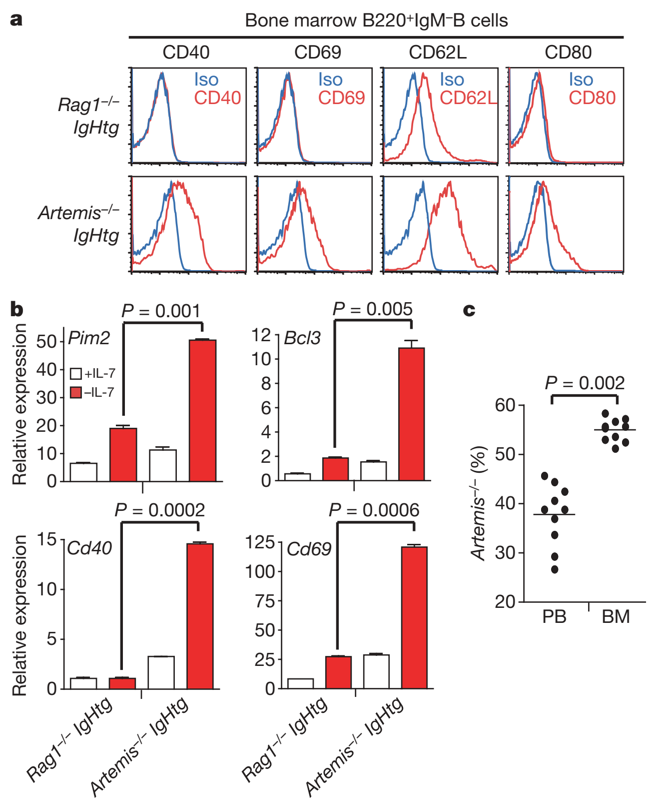


Figure 2 | Rag DSB-dependent gene expression changes in developing B cells *in vivo*. a, Flow cytometric analyses of CD40, CD69, CD80 and CD62L expression by B220^{+} IgM^{-} bone marrow B cells (primarily pre-B cells) from *Rag1*$^{-/-}$ *IgHtg* and *Artemis*$^{-/-}$ *IgHtg* mice (gated as shown in Supplementary Fig. 8). Histograms for specific antibodies (red) and isotype controls (Iso, blue) are shown. **b,** qRT–PCR of gene expression in bone marrow cultures derived from *Rag1*$^{-/-}$ *IgHtg* and *Artemis*$^{-/-}$ *IgHtg* mice before (+IL-7, white bar) and 48 h after (−IL-7, red bar) the removal of IL-7. Mean and s.d. from two experiments are shown; *P* values were calculated using a Student's one-tailed *t*-test. **c,** Percentage of *Artemis*$^{-/-}$ abl pre-B cells in the peripheral blood (PB) and bone marrow (BM) of Rag-deficient mice after co-injection of STI571-treated *Artemis*$^{-/-}$ and *Rag-2*$^{-/-}$ abl pre-B cells at a 1:1 ratio. The mean (line) of each set of mice analysed is indicated and *P* values were calculated using a two-tailed Wilcoxon matched-pairs test.

the genes that are regulated by Rag DSBs in abl pre-B cells (Fig. 2b, Supplementary Fig. 10 and data not shown)[18]. Together, these findings demonstrate that Rag DSBs activate a broad genetic program in developing pre-B cells. Notably, these DNA breaks also promote gene expression changes in pro-B and pro-T cells, suggesting that transcriptional programs are activated by Rag DSBs in all developing lymphocytes (Supplementary Fig. 11).

The genetic program induced by Rag DSBs includes genes that are active in many cellular processes. Several have functions that would be part of canonical DNA damage responses; for example, the anti-apoptotic proteins Pim2 and Bcl3 could antagonize pro-apoptotic pathways, such as those initiated by p53, promoting the survival of developing lymphocytes with Rag DSBs[19–21]. However, this program also includes genes that have no known role in canonical DNA damage responses; instead, these genes participate in diverse processes, many of which are important for lymphocyte development and function. These include common γ-chain (also known as Il2rg), SWAP70 (SWA-70), Notch 1, CD69, CD40, CD80 (B7.1) and CD62L (L-selectin). Notably, CD62L, SWAP70 and CD69 have known functions in lymphocyte homing and migration, suggesting that these gene expression changes could regulate the localization of developing lymphocytes with Rag DSBs[22–26]. Indeed, we observed a biased accumulation of *Artemis*$^{-/-}$ abl pre-B cells with Rag DSBs, as compared to *Rag2*$^{-/-}$ abl pre-B cells, in the bone marrow after co-injection of these cells into Rag-deficient mice (Fig. 2c and Supplementary Fig. 12). A similar bias was not observed in the peripheral blood of these mice (Fig. 2c). Moreover, inhibition of Atm signalling in *Artemis*$^{-/-}$ abl pre-B cells led to the diminished localization of these cells in the bone marrow (Supplementary Fig. 12c). Similar results were obtained from analyses of independently derived *Artemis*$^{-/-}$ and *Rag2*$^{-/-}$ abl pre-B cells (data not shown). Lymphocytes traverse different microenvironments as they develop in the thymus and bone marrow; our findings suggest that one of the functions of the Rag DSB-dependent genetic program may be to modulate this migration[27–30].

We have established that persistent Rag DSBs in NHEJ-deficient lymphocytes activate a diverse genetic program *in vitro* and *in vivo*. In wild-type lymphocytes, most Rag DSBs are repaired rapidly; thus, for this program to regulate processes that are generally important for development, it must also be initiated by these transient DSBs. In this regard, EMSA analyses did not show a marked increase in NF-κB activation in wild-type abl pre-B cells undergoing V(D)J recombination (Supplementary Fig. 13). However, this analysis might not be sensitive enough to detect the activation of NF-κB in response to transient Rag DSBs. In agreement with this notion, we found that NF-κB-dependent, Rag DSB-responsive genes are upregulated in an Atm-dependent manner in wild-type developing B cells undergoing V(D)J recombination (Fig. 3a and Supplementary Figs 9 and 14).

To determine directly whether transient Rag DSBs activate transcriptional pathways, we developed a sensitive single-cell assay for NF-κB activation. To this end, a retroviral reporter, pMSCV-NRE-GFP, that contains five tandem NF-κB responsive elements (NREs) driving expression of the green fluorescent protein (GFP) complementary DNA on the antisense strand, was generated (Supplementary Fig. 15a). pMSCV-NRE-GFP was introduced into *Rag2*$^{-/-}$ abl pre-B cells that contained the pMX-DEL^{CJ} retroviral V(D)J recombination substrate, hereafter referred to as *Rag2*$^{-/-}$ *NRE* cells (Fig. 3b and Supplementary Fig. 15)[14]. These cells do not express significant levels of GFP in response to STI571, but they do upregulate GFP in an Atm- and NF-κB-dependent manner in response to DSBs induced by ionizing radiation (Fig. 3b and Supplementary Fig. 16).

To determine whether transient Rag DSBs activate NF-κB, *Rag2*$^{-/-}$ *NRE* cells were transduced with a retrovirus encoding Rag2 (hereafter referred to as *Rag2*$^{-/-}$ *NRE R2* cells), which permits efficient V(D)J recombination after induction of Rag1 expression (Fig. 3c and data not shown). As STI571-treated cells are not dividing, GFP produced in response to transient Rag DSBs will persist, marking any cell that has activated NF-κB (Fig. 3b and Supplementary Fig. 15c). Notably, after 48 h of STI571 treatment, 53% of *Rag2*$^{-/-}$ *NRE R2* cells were GFP^{+} (Fig. 3b). This NF-κB activation was Atm-dependent, as an Atm inhibitor abrogated GFP expression (Fig. 3b). If NF-κB activation was stimulated by transient Rag DSBs, then there should be a higher level of completed rearrangements in the GFP^{+} cells. Indeed, PCR analyses demonstrated a

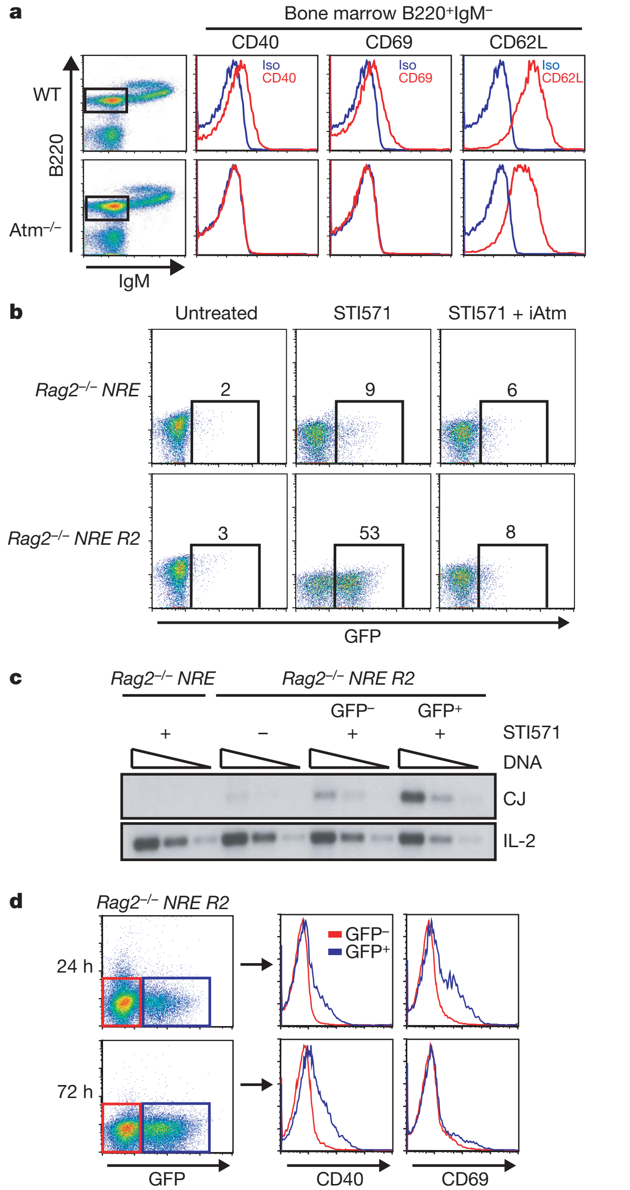


Figure 3 | NF-κB activation in response to transient Rag DSBs. **a**, Flow cytometric analyses of CD40, CD69 and CD62L expression by B220^{+}:IgM^{-} bone marrow B cells (gated cells in dot plot) from wild type (WT) and *Atm*$^{-/-}$ mice. Histograms for the specific antibodies (red) and isotype control (Iso, blue) are shown. **b**, Flow cytometric analysis of GFP expression in *Rag2*$^{-/-}$ *NRE* and *Rag2*$^{-/-}$ *NRE R2* cells treated with STI571 for 48 h in the presence or absence of the Atm inhibitor KU-55933 (iAtm). The percentage of GFP^{+} cells is indicated. Results are representative of three experiments. **c**, *Rag2*$^{-/-}$ *NRE* and *Rag2*$^{-/-}$ *NRE R2* abl pre-B cells were untreated (−) or treated with STI571 for 48 h (+), and GFP^{+} and GFP^{-} *Rag2*$^{-/-}$ *NRE R2* abl pre-B cells were isolated by flow cytometric cell sorting. Serial fourfold dilutions of genomic DNA from all cells were assayed for pMX-DEL^{CJ} rearrangement by PCR (see Supplementary Fig. 15b). Results are representative of two experiments. CJ, coding joint. The IL-2 gene PCR is shown as a DNA quantity control. **d**, Flow cytometric analysis of CD40 and CD69 protein expression on GFP^{+} (blue histograms) and GFP^{-} (red histograms) *Rag-2*$^{-/-}$ *NRE R2* cells treated with STI571 for 24 or 72 h. Results are representative of two experiments.

fivefold higher level of pMX-DEL^{CJ} rearrangement (coding joint, CJ, formation) in the GFP^{+} cells as compared to GFP^{-} cells (Fig. 3c). Moreover, the GFP^{+} cells showed increased expression of genes regulated by Rag DSBs, including *Pim2*, *Bcl3*, *Cd40* and *Cd69* (Fig. 3d and Supplementary Fig. 17). CD69 expression in the GFP^{+} population was transient, similar to that observed in response to persistent Rag DSBs (Fig. 3d and Supplementary Figs 4a and 5b). Together, these data demonstrate that transient Rag DSBs generated during normal V(D)J recombination activate DNA DSB responses, including Atm and NF-κB, which lead to the initiation of a diverse genetic program in all developing lymphocytes.

As physiological Rag DSBs activate a broad genetic program, we reasoned that genotoxic DSBs may also activate components of this program in developing lymphocytes. Indeed, treatment of STI571-treated *Rag2*$^{-/-}$ abl pre-B cells with either ionizing radiation or etoposide—which both generate DNA DSBs—led to the Atm-dependent induction of many genes that are also induced by Rag DSBs (Fig. 4a, b and Supplementary Fig. 18). Furthermore, treatment of bone marrow from *Rag1*$^{-/-}$ *IgHtg* mice with either ionizing radiation or etoposide led to a B-cell-specific (B220^{+}), Atm-dependent induction of CD69 expression (Fig. 4c). These findings demonstrate that, in developing B cells, genotoxic DSBs can induce lymphocyte-specific gene expression changes that are normally induced by Rag DSBs.

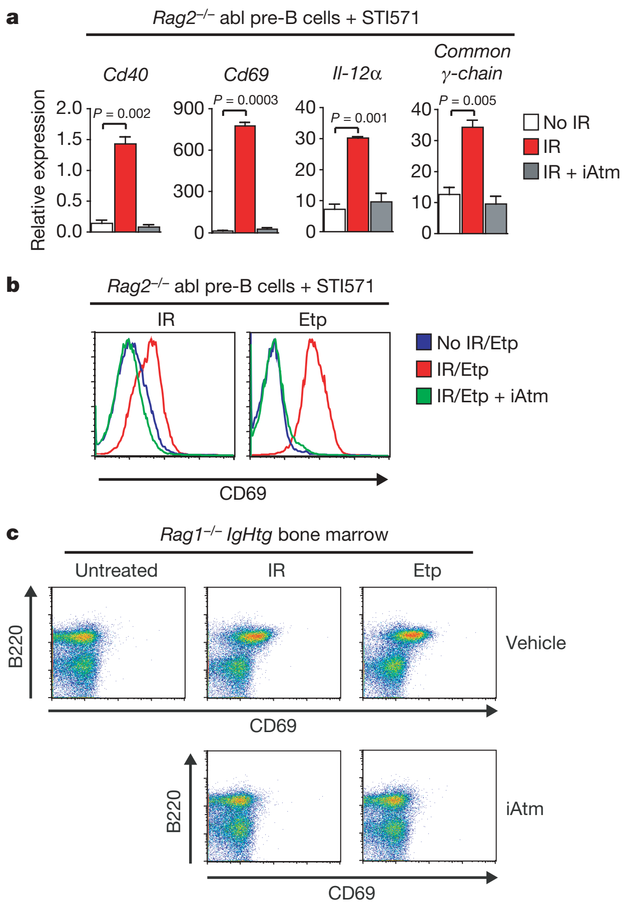


Figure 4 | Genotoxic DSBs promote changes in expression of lymphocyte-specific genes. **a**, qRT–PCR analysis of gene expression in *Rag2*$^{-/-}$ abl pre-B cells 2 h after receiving 0 Gy (white bars) or 4 Gy of ionizing radiation (IR) in the presence (grey bars) or absence (red bars) of the Atm inhibitor KU-55933 (iAtm). All cells were treated with STI571 for 24 h before ionizing radiation. Results are the mean and s.d. from two experiments; *P* values were calculated using a Student's one-tailed *t*-test. **b**, Flow cytometric analysis of CD69 expression on STI571-treated *Rag2*$^{-/-}$ abl pre-B cells treated with ionizing radiation as described in **a** or 2 h after being treated with 5 μM etoposide (Etp). Results are representative of three experiments. **c**, Flow cytometric analysis of B220 and CD69 expression on bone marrow cells from *Rag1*$^{-/-}$ *IgHtg* mice 2 h after irradiation (0.5 Gy) or etoposide treatment (5 μM) in the presence or absence of the Atm inhibitor KU-55933. Results are representative of two experiments.

We have shown that physiological DNA DSBs generated during antigen receptor gene assembly activate a broad genetic program in developing lymphocytes. Notably, this program is initiated by transient Rag DSBs that are rapidly repaired during normal V(D)J recombination in all developing lymphocytes. Thus, we propose that Rag DSBs provide cellular cues that regulate processes—such as migration and homing—that are integral components of normal lymphocyte development. As Atm is required to activate a portion of this Rag DSB-dependent genetic program, we expect that the defects in the expression of these genes contribute to the lymphopenia observed in ataxia-telangiectasia.

Our findings establish that DNA DSBs generated during physiological processes can regulate the cell-type-specific expression of genes that participate in functions not directly involved in maintaining genomic stability or suppressing malignant transformation. We speculate that DNA DSBs generated in other physiological settings, such as immunoglobulin class-switch recombination, meiosis and DNA replication, may have similar effects. Furthermore, genotoxic DSBs generated by agents such as ionizing radiation and chemotherapeutic drugs could corrupt normal cellular processes through the aberrant activation of these cell-type-specific genetic programs. Thus, the effect of DNA DSB responses on cell physiology and function is probably more pervasive than previously appreciated.

METHODS SUMMARY

Standard methodologies used for cell culture, retroviral vector generation, western blotting, Southern blotting, EMSA, gene expression profiling, PCR and flow cytometry are described in detail in the Methods. All v-abl-transformed pre-B cell lines were generated from mice harbouring the *EμBcl-2* transgene as previously described[14]. Flow cytometric analyses were performed on a FACSCaliber (BD Biosciences) with the normalized geometric mean fluorescence intensity calculated as the difference between the geometric mean fluorescence for the specific antibody and the isotype control. Flow cytometric cell sorting (FACS) was performed using a FACSVantage (BD Biosciences). To generate pMSCV-NRE-GFP, the NF-κB responsive elements and TATA box from pNF-κB-Luc (Stratagene) were amplified using the following oligonucleotides: pNRE-F, CCAAACTCATCAATGTATCTTATCATG and pNRE-R, TACCAACAGTACCGGAATGC. This PCR product was cloned upstream of a cDNA encoding enhanced GFP (Clontech) on the bottom strand of pMSCV containing an IRES upstream of the Thy1.2 cDNA. Gene expression profiling was performed using the Affymetrix 430 v2.0 mouse genome microarray. To assay abl pre-B cell localization, cells were treated for 48 h with STI571, labelled with 0.1 mM CFSE (Invitrogen) or 5 mM SNARF (Invitrogen) at 37 °C for 15 min in PBS at 2–5 × 10^6 ml^{-1}, washed three times, and resuspended in DMEM at a concentration of 80 × 10^6 ml^{-1}. Labelled cells were then mixed in a 1:1 ratio, and a total of 40 × 10^6 cells were injected into the tail vein of each Rag-deficient mouse. Bone marrow and peripheral blood was collected 30–60 min after injection and analysed by FACS for presence of CFSE- and SNARF-labelled abl pre-B cells.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 April; accepted 4 September 2008.
Published online 12 October 2008.

1. Shiloh, Y. ATM and related protein kinases: safeguarding genome integrity. *Nature Rev. Cancer* 3, 155–168 (2003).
2. Rouse, J. & Jackson, S. P. Interfaces between the detection, signaling, and repair of DNA damage. *Science* 297, 547–551 (2002).
3. Zhou, B. B. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* 408, 433–439 (2000).
4. Innes, C. L. *et al.* ATM requirement in gene expression responses to ionizing radiation in human lymphoblasts and fibroblasts. *Mol. Cancer Res.* 4, 197–207 (2006).
5. Rashi-Elkeles, S. *et al.* Parallel induction of ATM-dependent pro- and antiapoptotic signals in response to ionizing radiation in murine lymphoid tissue. *Oncogene* 25, 1584–1592 (2006).
6. Tonegawa, S. Somatic generation of antibody diversity. *Nature* 302, 575–581 (1983).

7. Fugmann, S. D. *et al.* The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu. Rev. Immunol.* **18,** 495–527 (2000).
8. Rooney, S., Chaudhuri, J. & Alt, F. W. The role of the non-homologous end-joining pathway in lymphocyte development. *Immunol. Rev.* **200,** 115–131 (2004).
9. Meek, D. W. The p53 response to DNA damage. *DNA Repair (Amst.)* **3,** 1049–1056 (2004).
10. Guidos, C. J. *et al.* V(D)J recombination activates a p53-dependent DNA damage checkpoint in scid lymphocyte precursors. *Genes Dev.* **10,** 2038–2054 (1996).
11. Perkins, N. D. Integrating cell-signalling pathways with NF-κB and IKK function. *Nature Rev. Mol. Cell Biol.* **8,** 49–62 (2007).
12. Huang, T. T., Wuerzberger-Davis, S. M., Wu, Z. H. & Miyamoto, S. Sequential modification of NEMO/IKKγ by SUMO-1 and ubiquitin mediates NF-κB activation by genotoxic stress. *Cell* **115,** 565–576 (2003).
13. Gerondakis, S. *et al.* Unravelling the complexities of the NF-κB signalling pathway using mouse knockout and transgenic models. *Oncogene* **25,** 6781–6799 (2006).
14. Bredemeyer, A. L. *et al.* ATM stabilizes DNA double-strand-break complexes during V(D)J recombination. *Nature* **442,** 466–470 (2006).
15. Muljo, S. A. & Schlissel, M. S. A small molecule Abl kinase inhibitor induces differentiation of Abelson virus-transformed pre-B cell lines. *Nature Immunol.* **4,** 31–37 (2003).
16. May, M. J. *et al.* Selective inhibition of NF-κB activation by a peptide that blocks the interaction of NEMO with the IκB kinase complex. *Science* **289,** 1550–1554 (2000).
17. Wu, Z. H., Shi, Y., Tibbetts, R. S. & Miyamoto, S. Molecular linkage between the kinase ATM and NF-κB signaling in response to genotoxic stimuli. *Science* **311,** 1141–1146 (2006).
18. Grawunder, U. *et al.* Expression of DNA-dependent protein kinase holoenzyme upon induction of lymphocyte differentiation and V(D)J recombination. *Eur. J. Biochem.* **241,** 931–940 (1996).
19. Kashatus, D., Cogswell, P. & Baldwin, A. S. Expression of the Bcl-3 proto-oncogene suppresses p53 activation. *Genes Dev.* **20,** 225–235 (2006).
20. Amaravadi, R. & Thompson, C. B. The survival kinases Akt and Pim as potential pharmacological targets. *J. Clin. Invest.* **115,** 2618–2624 (2005).
21. Hayden, M. S. & Ghosh, S. Signaling to NF-κB. *Genes Dev.* **18,** 2195–2224 (2004).
22. Rosen, S. D. Ligands for L-selectin: homing, inflammation, and beyond. *Annu. Rev. Immunol.* **22,** 129–156 (2004).
23. Feng, C. *et al.* A potential role for CD69 in thymocyte emigration. *Int. Immunol.* **14,** 535–544 (2002).
24. Matloubian, M. *et al.* Lymphocyte egress from thymus and peripheral lymphoid organs is dependent on S1P receptor 1. *Nature* **427,** 355–360 (2004).
25. Shiow, L. R. *et al.* CD69 acts downstream of interferon-α/β to inhibit S1P1 and lymphocyte egress from lymphoid organs. *Nature* **440,** 540–544 (2006).
26. Pearce, G. *et al.* Signaling protein SWAP-70 is required for efficient B cell homing to lymphoid organs. *Nature Immunol.* **7,** 827–834 (2006).
27. Ladi, E., Yin, X., Chtanova, T. & Robey, E. A. Thymic microenvironments for T cell differentiation and selection. *Nature Immunol.* **7,** 338–343 (2006).
28. Johnson, K. *et al.* Regulation of immunoglobulin light-chain recombination by the transcription factor IRF-4 and the attenuation of interleukin-7 signaling. *Immunity* **28,** 335–345 (2008).
29. Tokoyoda, K. *et al.* Cellular niches controlling B lymphocyte behavior within bone marrow during development. *Immunity* **20,** 707–718 (2004).
30. Nagasawa, T. Microenvironmental niches in the bone marrow required for B-cell development. *Nature Rev. Immunol.* **6,** 107–116 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Bednarski, B. Van Honten and M. Diaz for critical review of the manuscript, F. W. Alt for providing us with the *Artemis*$^{-/-}$ mice and D. Ballard for providing us with the IκBα-ΔN construct. This research is supported by the National Institutes of Health (NIH, grant AI47829) and the Washington University Department of Pathology and Immunology (to B.P.S.); the Department of Pathology and Center for Childhood Cancer Research of the Children's Hospital of Philadelphia, and the Abramson Family Cancer Research Institute (to C.H.B.); and the intramural research program of the NIH, National Institute of Environmental Health Sciences (to R.S.P.). B.P.S. is a recipient of a Research Scholar Award from the American Cancer Society. C.H.B. is a Pew Scholar in the Biomedical Sciences. A.L.B. is supported by a post-doctoral training grant from the NIH.

Author Information The microarray gene expression data have been deposited in NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE9024. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.P.S. (Sleckman@immunology.wustl.edu).

METHODS

Mice. The $Atm^{-/-}$, Scid, $Ku70^{-/-}$, $Artemis^{-/-}$, VH147 IgH transgenic and EμBcl-2 transgenic mice have been described previously[14,31–33]. Mice were bred and maintained under specific pathogen-free conditions at the Washington University School of Medicine and were handled in accordance to the guidelines set forth by the Division of Comparative Medicine of Washington University.

Cell culture. Three independently derived $Rag2^{-/-}$ (R.1, R.2 and R.3), three independently derived $Artemis^{-/-}$ (A.1, A.2 and A.3), two independently derived $Artemis^{-/-}$ $Atm^{-/-}$ (AA.1 and AA.2) and individual Scid and $Ku70^{-/-}$ v-abl-transformed pre-B cell lines containing the EμBcl-2 transgene were generated. The three $Artemis^{-/-}$ *IκBα-ΔN* cells (A.3ΔN1, A.3ΔN2 and A.3ΔN3) were generated by standard transfection of the A.3 cells with a cDNA encoding *IκBα-ΔN*[34]. Expression of the IκBα-ΔN protein in these cells was confirmed by western blotting. STI571 treatments were carried out using 3 μM STI571 as previously described[14]. The KU-55933 Atm inhibitor (Sigma) was used at 15 μM, the BAY-11-7085 inhibitor (Calbiochem) was used at 20 μM, and etoposide was used at 5 μM. Irradiation was carried out with a Cs^{137} source, at doses of either 0.5 or 4 Gy.

Southern blotting. Southern blot analysis was performed as described previously[14,35].

Western blotting and EMSA. Western blotting was carried out as described previously, using an antibody to Pim2 (Santa Cruz, 1D12)[36]. The secondary reagent was horseradish peroxidase (HRP)-conjugated goat anti-mouse IgG (Zymed). EMSAs were run as described previously and analysed using a Li-Cor Odyssey Infrared scanner[37]. Supershifts were performed using anti-p50 and anti-p65 antibodies (Santa Cruz, sc-114 X and sc-372 X, respectively).

Gene expression profiling. All abl pre-B cells were collected 48 h after treatment with STI571. RNA was isolated using Qiagen RNeasy technology following the manufacturer's instructions. Gene expression analysis was conducted using Affymetrix Mouse Genome 2.0 GeneChip arrays (Mouse 430 v2, Affymetrix). One microgram of total RNA was amplified as directed in the Affymetrix One-Cycle cDNA Synthesis protocol. Fifteen micrograms of amplified biotin-complementary-RNAs were fragmented and hybridized to each array for 16 h at 45 °C in a rotating hybridization oven using the Affymetrix Eukaryotic Target Hybridization Controls and protocol. Array slides were stained with streptavidin and phycoerythrin using a double-antibody staining procedure, and then washed using the EukGE-WS2v5 protocol of the Affymetrix Fluidics Station FS450 for antibody amplification. Arrays were scanned in an Affymetrix Scanner 3000 and data was obtained using the GeneChip Operating Software (Version 1.2.0.037). Data pre-processing, normalization and error modelling was performed with the Rosetta Resolver system (Version 6.0)[38].

To identify differentially expressed genes between the $Rag2^{-/-}$ and the $Artemis^{-/-}$ abl pre-B cells, an error-weighted analysis of variance (ANOVA) using the Benjamini–Hochberg false discovery rate was performed using Rosetta Resolver. At $P \leq 0.05$, this analysis yielded 1,578 probe sets. Fold changes were generated in Resolver, based on the ratio of each individual cell line relative to the average of the three $Rag2^{-/-}$ cells with 100-fold being the maximum change that Resolver can report. Of the 1,578 probes found using the ANOVA analysis, only those with a fold-change ≥2.0 (not including anti-correlated genes) in all ratios for a given genotype were considered for further analysis. A Venn diagram of the resulting ANOVA and fold-change data sets for each genotype comparison to $Rag2^{-/-}$ was then used to determine which expression changes were dependent on Atm and NF-κB.

For gene expression profiling of primary developing B cells, $B220^{+}$ cells were isolated from $Rag1^{-/-}$ *IgHtg* and $Art^{-/-}$ *IgHtg* bone marrow using Mouse Pan B (B220) Dynabeads from Invitrogen. RNA was isolated using the Qiagen RNeasy Mini Kit. Microarrays were treated as described above for the abl pre-B cells. For the probe sets that were identified as being differentially regulated in response to Rag DSBs in the abl pre-B cells, we compared the average expression in the two $Rag1^{-/-}$ *IgHtg* samples to each of the two $Art^{-/-}$ *IgHtg* samples. The probe sets that were not anti-correlated and changed ≥1.5-fold in the average of these comparisons are included in Supplementary Table 4.

Flow cytometry. Flow cytometric analyses were performed using fluorescein isothiocyanate (FITC)-conjugated anti-CD45R/B220, allophycocyanin (APC)-conjugated anti-IgM, phycoerythrin (PE)-conjugated anti-CD40, PE-conjugated anti-CD62L, PE-conjugated anti-CD69, and the appropriate isotype controls (BD Biosciences).

Analysis of pMX-DEL^{CJ} rearrangement. PCR was carried out as described previously on fourfold dilutions of genomic DNA[14]. IL-2 PCR primers were IMR42 and IMR43, and the product was probed with IMR042-2, all of which have been described previously[1]. Primer sequences for pMX-DEL^{CJ} are as follows: pA, CACAGGATCCCACGAAGTCTTGAGACCT; pB, ATCTGGATCC-GTGCCGCCTTTGCAGGTGTATC; pD, AGACGGCAATATGGTGGA.

qRT–PCR gene expression analysis. qRT–PCR was performed using a Stratagene Mx3000P real-time PCR machine and Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen). PCRs were carried out using the following program: 95 °C, 30 s; 55 °C, 60 s; 72 °C, 30 s; 40 cycles. Relative expression is calculated as the difference between expression of β-actin (C1) and expression of the gene of interest (C2), using the equation relative expression $= 2^{-(C1-C2)}$. C is the PCR cycle in which the product detection curve becomes linear. Primers are listed in Supplementary Table 5.

31. Rooney, S. *et al.* Leaky Scid phenotype associated with defective V(D)J coding end processing in Artemis-deficient mice. *Mol. Cell* **10**, 1379–1390 (2002).
32. Gu, Y. *et al.* Growth retardation and leaky SCID phenotype of Ku70-deficient mice. *Immunity* **7**, 653–665 (1997).
33. Mandik-Nayak, L., Racz, J., Sleckman, B. P. & Allen, P. M. Autoreactive marginal zone B cells are spontaneously activated but lymph node B cells require T cell help. *J. Exp. Med.* **203**, 1985–1998 (2006).
34. Brockman, J. A. *et al.* Coupling of a signal response domain in IκBα to multiple pathways for NF-κB activation. *Mol. Cell. Biol.* **15**, 2809–2818 (1995).
35. Ramsden, D. A. & Gellert, M. Formation and resolution of double-strand break intermediates in V(D)J rearrangement. *Genes Dev.* **9**, 2409–2420 (1995).
36. Khor, B. *et al.* Proteasome activator PA200 is required for normal spermatogenesis. *Mol. Cell. Biol.* **26**, 2999–3007 (2006).
37. Weaver, B. K., Bohn, E., Judd, B. A., Gil, M. P. & Schreiber, R. D. ABIN-3: a molecular basis for species divergence in interleukin-10-induced anti-inflammatory actions. *Mol. Cell. Biol.* **27**, 4603–4616 (2007).
38. Weng, L. *et al.* Rosetta error model for gene expression analysis. *Bioinformatics* **22**, 1111–1121 (2006).

ERRATUM

doi:10.1038/nature07622

Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer

Ryohei Sekido & Robin Lovell-Badge

Nature 453, 930–934 (2008)

The Supplementary Methods for this Letter should have been uploaded at the time of publication. This oversight has now been rectified.

RETRACTION

doi:10.1038/nature07645

The RNA-binding protein FCA is an abscisic acid receptor

Fawzi A. Razem, Ashraf El-Kereamy, Suzanne R. Abrams & Robert D. Hill

Nature 439, 290–294 (2006)

Portions of the work repeated with respect to abscisic acid (ABA) binding have revealed errors in the calculations associated with Fig. 1, with the result that the molar ratio of ABA bound to FCA is substantially lower than claimed. There are also difficulties with the data in Fig. 2a, b that arose from the preparation of FY. We conclude that there is no effect of ABA on the FCA–FY interaction, and therefore requested to retract this paper on 14 July 2008. See the Brief Communication Arising in this issue[1].

1. Risk, J. M., Macknight, R. C. & Day, C. L. FCA does not bind abscisic acid. *Nature* doi:10.1038/nature07646 (this issue).

ERRATUM

doi:10.1038/nature07622

Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer

Ryohei Sekido & Robin Lovell-Badge

Nature 453, 930–934 (2008)

The Supplementary Methods for this Letter should have been uploaded at the time of publication. This oversight has now been rectified.

RETRACTION

doi:10.1038/nature07645

The RNA-binding protein FCA is an abscisic acid receptor

Fawzi A. Razem, Ashraf El-Kereamy, Suzanne R. Abrams & Robert D. Hill

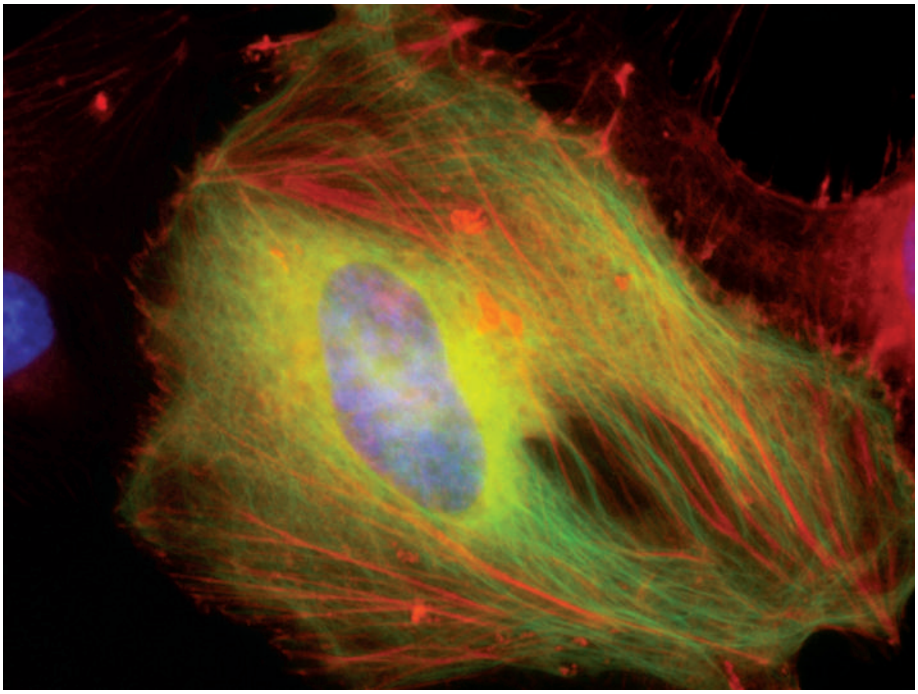
Nature 439, 290–294 (2006)

Portions of the work repeated with respect to abscisic acid (ABA) binding have revealed errors in the calculations associated with Fig. 1, with the result that the molar ratio of ABA bound to FCA is substantially lower than claimed. There are also difficulties with the data in Fig. 2a, b that arose from the preparation of FY. We conclude that there is no effect of ABA on the FCA–FY interaction, and therefore requested to retract this paper on 14 July 2008. See the Brief Communication Arising in this issue[1].

1. Risk, J. M., Macknight, R. C. & Day, C. L. FCA does not bind abscisic acid. *Nature* doi:10.1038/nature07646 (this issue).

New ways to see a smaller world

Advances in imaging are allowing researchers to gain better insights into the function of tissues, cells and even individual molecules. Nathan Blow examines the latest technologies lighting the way.



INVITROGEN

Fluorescent probes and dyes can be used to image the inner world of the cell.

Competition is often a stimulus for change. But a competition can also be a mirror of a technology's evolution. Take the annual Nikon International Small World Competition and the Olympus BioScapes International Digital Imaging Competition, for example: these events give scientists from around the world the opportunity to submit their best photographed microscopy images for judging by panels of experts.

This year, the winner of the Small World contest, announced in mid-October, was Michael Stringer from Westcliff-on-Sea, UK, for a multicolour, dark-field image of marine diatoms. The winner of the 2008 BioScapes contest was Spike Walker of Penkridge, UK, using Rheinberg illumination of a fairy fly wasp. Entrants in these competitions are using everything from conventional stereomicroscopes to the latest in laser-scanning confocal instrumentation, in some cases with different coloured fluorophores and the newest three-dimensional rendering software, to obtain images ranging from Stringer's diatoms, to a growing bundle of carbon nanotubes, to colourful pictures of individually labelled neurons in a mouse brain.

According to Eric Flem, director of the Nikon competition — now in its 34th year — the size and scope of the contest has grown exponentially in recent years, something he attributes to advances in the underlying imaging technology. "Imaging has become so much easier to do and so much more powerful — it is the norm rather than the exception in most scientific fields these days," he says.

Probing questions

Researchers often use a little beacon of light, a fluorophore, to see the inside world of the cell. Although a great number and range of fluorescent proteins are currently available for this purpose, improving these to meet the expanding needs of cell biologists can be quite a challenge for developers.

"Often improvements in one characteristic of a fluorescent protein can be reached," says Ilya Kelmanson, head of the product department at fluorescent-protein developer Evrogen in Moscow, "but usually at the expense of another important parameter." An example of this is monomerization of fluorescent proteins, which helps in tagging proteins within the cell, but will generally result in a decrease in the fluorescence intensity. To deal with this, Evrogen often maintains different lines of fluorescent proteins: bright dimeric probes (made of two monomers) used for cell-labelling applications, and less-bright monomeric versions for tagging purposes. Kelmanson says that Evrogen is developing a third line of fluorescent proteins that will consist of tandem versions of dimeric fluorescent proteins that have stronger fluorescence than their monomeric counterparts, and good performance and brightness as fusion partners with other proteins.

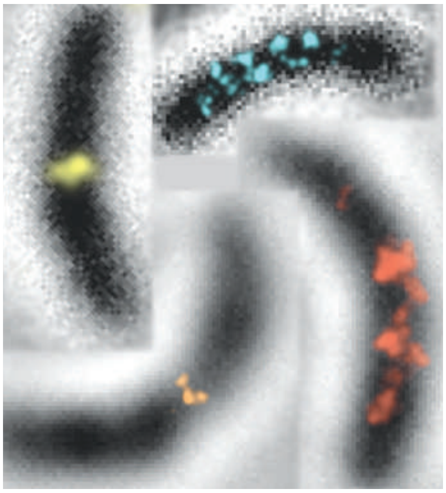
There are many types of fluorescent proteins that emit in the middle range of the light spectrum — green to yellow. But it is a different story at the far ends of the spectrum. "We still do not have good fluorescent proteins emitting in the near-infrared range," says Kelmanson. Probes in this range could be very useful for whole-body imaging applications and multicolour labelling experiments, he says.

"People are moving away from the blue and the green excitation wavelengths and into the red and near-infrared," says Martin Hoppe, a market manager at microscope developer Leica Microsystems in Wetzlar, Germany. He says that using fluorescent proteins with longer excitation wavelengths helps researchers when it comes to live-cell imaging applications — an area of keen interest for most cell biologists (see 'Light activated') — because these wavelengths tend to be less damaging to cells. Newer dyes have been optimized for near-infrared excitation. One of these, mCherry, one of the 'fruit fluorescent proteins' developed in the lab of Roger Tsien at the University of California, San Diego — who shared the 2008 Nobel Prize in Chemistry for the discovery and development of green fluorescent protein — is now offered by Clontech in Mountain View, California, along with several fluorescent proteins derived from reef corals.

Developing fluorescent proteins at longer wavelengths, particularly for fluorescence resonance energy transfer (FRET) applications, which are used to measure interactions between two proteins *in vivo*, "has been a challenge", says George Hanson, a principal scientist working on cell signalling at Life Technologies (formerly Invitrogen) in Eugene, Oregon. Although a challenge, there are potential benefits — FRET efficiency grows and background fluorescence decreases when using longer wavelength fluorophores, according to Kelmanson.

Still, the development of fluorescent proteins that excite at longer wavelengths is only part of the issue for those interested in applying these fluorophores to their FRET studies. Although there are many cyanine to yellow FRET pairs available along with filters and instrumentation preset for reading their interactions, the use of more red-shifted fluorescent protein pairs might take time as optimal parameters and appropriate microscopy hardware and software will need to be developed.

At Life Technologies, Hanson has been working



J. BITEEN

Fluorescence switching is key to single-molecule imaging methods such as PALM and STORM.

to advance the use of FRET in high-throughput screening applications. Here, the challenge of obtaining a strong FRET signal is crucial to overcoming any background fluorescence. Because the simple approach of making a brighter fluorescent protein can be time-consuming, another approach, being used by a number of developers, is to take advantage of the 'states' — the differences between the excitation and emission — of a fluorescent protein. Time resolved FRET (TR-FRET) uses long-lived fluorophores that exhibit a time delay between their excitation and emission, minimizing potential interference from background fluorescence. Several companies now offer TR-FRET platforms for cellular interaction assays, including Life Technologies, Cisbio in Bedford, Massachusetts, Covalys of Witterswill, Switzerland, and PerkinElmer located in Waltham, Massachusetts.

Quantum of happiness

Another route to fluorescence is provided by quantum dots, which "present better photostability than organic dyes and allow for increased multiplexing capability", says Stephen Chamberlain, manager of business development at Life Technologies.

Quantum dots are composed of a semiconductor core shell surrounded by an exterior coating. What makes them unique as fluorophores is their ability to be excited by a single wavelength of light — the emission from a quantum dot is dictated by the size of its core shell. As the size of the core shell can be varied to obtain different light-emission spectra, quantum dots present a particularly interesting solution for researchers interested in experiments using multiple coloured fluorophores.

But unlike fluorescent proteins, quantum dots are not genetically encodable, which has been a stumbling block to more widespread usage. "It is hard to get them into the cell," says Alice Ting, a chemist at the Massachusetts Institute of Technology in Cambridge, who has been working with quantum dots for several years. "And then once inside, how do you target them to a specific protein?"

Most researchers and developers are looking towards the outer coating for targeting. At Life Technologies, says Chamberlain, they place a polymer coating around the core shell of their Qdot Nanocrystals along with an additional layer of polyethylene glycol on top of the polymer coating to reduce the sticking together of the Qdots, and to provide a surface for the attachment of antibodies and other molecules for targeting.

Despite the challenges of targeting, quantum dots are now being tried in a number of imaging applications. Chamberlain says that like their organic-dye counterparts 20 years ago, researchers are still in the process of developing robust application protocols for quantum dots. Although quantum dots can be used for FRET applications, much like fluorescent proteins, they tend to make better donors in a FRET interaction because they can be excited by light that is blue-shifted relative to their emission. They are being used for *in vivo* imaging applications, such as looking at blood vasculature, because of their high residence times and ability to excite in the near-infrared region of the spectrum.

Ting says using quantum dots for single-molecule imaging is a much more straightforward application at the moment, although the issues of delivery and targeting still exist, and several groups are now taking advantage of the greater brightness of quantum dots to track the movement of single molecules within cells.

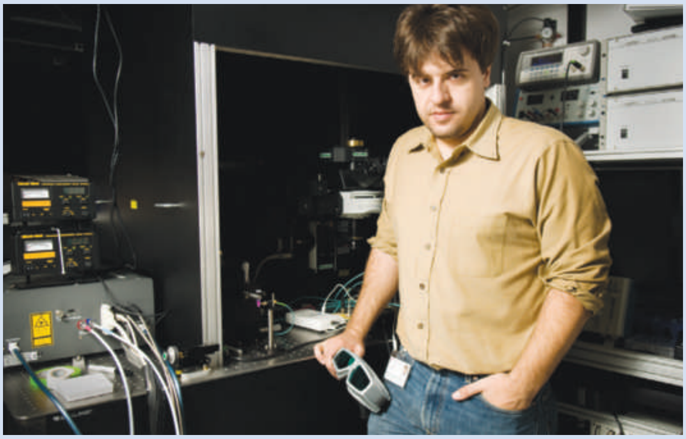
Microscopes respond

As the types and range of applications of fluorescent probes and quantum dots continue to grow, developers of confocal microscopes have been working hard to keep instrumentation

LIGHT ACTIVATED

When scientists first discovered light-absorbing molecules — in microorganisms — that could control the flow of ions into a cell, many researchers became interested in comparing how these proteins differed from their visual pigment counterparts. But Karl Deisseroth at Stanford University in California saw something else in these light-activated channels — a potential tool to dissect the function of the brain.

Deisseroth — along with Edward Boyden, his postdoctoral fellow at the time, now at the Massachusetts Institute of Technology in Cambridge, and his graduate student Feng Zhang — collaborated with microbiologists to hunt for these light-activated channels from various species. "I come from a synaptic physiology background, so we just sequentially tried these and other proteins to find the ones that worked the best," says Deisseroth. It was a collaboration with Georg Nagel and Ernst Bamberg at the Max Planck Institute of Biophysics in Frankfurt, Germany, who supplied a clone of a gene called channelrhodopsin-2 from *Chlamydomonas reinhardtii* to Deisseroth that opened the floodgates.



K. DEISSEROTH

Karl Deisseroth develops optogenetic methods for controlling neural activation.

Channelrhodopsin-2 is a gated light-sensitive cation channel that uses a molecule of all-*trans* retinal to absorb photons. When Deisseroth, Boyden and Zhang expressed channelrhodopsin-2 in hippocampal neurons in the mouse brain then shone blue light on the region, they found the cells with channelrhodopsin-2 responded to the light stimulation, opening the channel and initiating the flow of ions, which resulted in an action potential in those neurons[6]. Deisseroth says that one of the most pleasant surprises to emerge from that first series of experiments turned out to be the precision of the system. "Even though these molecules are not designed to generate action potentials and work on this time scale in neurons," he says, "it turns out that they can."

This result led to the start of a new field, coined 'optogenetics' by Deisseroth in 2006, where researchers are combining optics with genetics to explore the workings of neural circuits. In the years since the first description of channelrhodopsin-2, Deisseroth's lab has gone on to advance the system and develop new probes. The team identified and developed two more light-activated

up to the challenge. "In general, researchers are still requesting the same three things: higher speed, better signal-to-noise and lower toxicity," says Michael Davis, a confocal product manager from Nikon Instruments in Melville, New York. But he is quick to add that there are several emerging fluorescent-protein techniques that have required more sophisticated confocal technology.

Davis says more researchers are now using photoactivatable and photoconvertible proteins in their research. The idea here is to use a specific wavelength of light to excite the fluorescent protein, which can then be turned off, in the case of photoactivatable proteins, or even converted to another colour, in the case of photoconvertible proteins, with a different wavelength of light. Evrogen now offers both kindling red fluorescent protein (KFP-Red), which can be either reversibly or irreversibly activated for either short or long-term protein tracking in a cell, and PS-CFP2, which is a photoconvertible protein that will switch from cyan to green in response to light activation. And MBL International in Woburn, Massachusetts, offers the photoconvertible proteins CoralHue Kaede and CoralHue Kikume, which convert from green to red in response to light activation.

The development of these fluorescent proteins has led to dual scanning systems for confocal microscopes, allowing for the simultaneous imaging and activation of photoactivatable and photoconvertible proteins, along with new spectral imaging technologies that can accurately assay a broad range of possible spectral emissions.

Another trend in the development of confocal instrumentation is the emergence of resonance scanners. "People would like to see the resolution they get out of a point scanner, but with the speed of a field scanner," says Davis. Several developers hope that resonance scanning will bridge this divide. With a resonance-scanning confocal system, the scanner mirrors are moved at extremely high frequencies, permitting much higher frame rates and minimizing the time the scanner dwells at any particular point in the sample, allowing researchers to study more dynamic events in a cell while reducing the effects of phototoxicity on it. Both Nikon and Leica Microsystems have introduced resonance-scanning confocal systems in recent years.

S. HELL

Stefan Hell developed the STED imaging approach.

Small switches

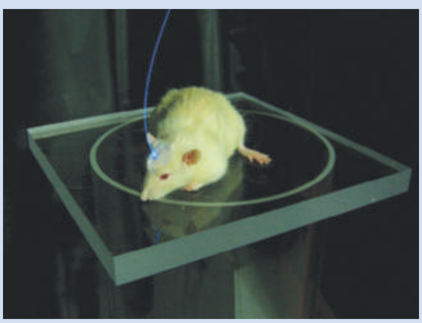
Although dual imaging and faster scanning are enhancing fluorescence imaging, the next step in improving the optical resolution of microscopes might require all developers to look more closely at the properties of those fluorescent proteins. "Optics alone can only provide resolution to a certain limit," says Davis.

The ability of a fluorescent protein or a dye to switch states is also at the heart of an emerging field in microscopy. "Fluorescence switching is the common basis to all super-resolution methods that have been devised so far," says Stefan Hell, director of the Max Planck Institute for Biophysical Chemistry in Gottingen, Germany, who created a super-resolution imaging method known as stimulated emission depletion (STED) 'nanoscopy'.

Before super-resolution imaging, the resolution limit of optical microscopy was thought to be determined by the diffraction barrier of light. For example, in a high-resolution confocal microscope, a beam of light is focused down to a spot on the focal plane of about 200 nanometres in size. Because all molecules within that spot are excited in parallel, the spatial resolution is limited to about this value. Hell's approach with STED was to use the transition states of fluorophores, or their photoswitching capabilities, to effectively reduce the size of that spot — potentially all the way down to a single molecule. In a STED microscope, the focal spot is overlapped with a ring of light that switches off fluorescence everywhere in the spot except at the centre of the ring. The outcome is a much smaller

molecules for optogenetic control of neurons. These were an inhibitor called halorhodopsin, and a *Volvox* channelrhodopsin that can also initiate neuronal activity, but is more than 100 nanometres red-shifted from the peak of the *Chlamydomonas* channelrhodopsin, allowing separable channels of optical control. They also developed various targeting strategies to get channels expressed in specific neuronal cell types and populations, and fibreoptic/laser-diode hardware for adaptation to mammals.

The growing number of researchers interested in applying optogenetic approaches to their particular research questions has led to an unexpected effort within Deisseroth's lab. "Technology distribution and helping other researchers has been another big part of what we are doing," says Deisseroth. His lab has supplied clones of the channels to more than 300 other labs around the globe for use in organisms ranging from mice to fruitflies. Although the channels have worked in every species tried thus far, in some cases, minor modifications have been required. Although there is sufficient all-*trans* retinal in the brains of mice for the channels to function properly, invertebrates must be supplemented with all-*trans* retinal through their diet to achieve channel activation.

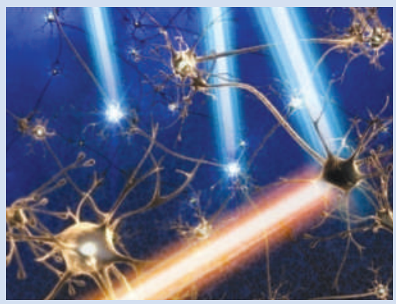


K. DEISSEROTH

Mouse brain neurons can be stimulated by light from fibre-optic cables attached to the animal.

As the basic use of these light-activated channels for studying brain function and circuitry gains more traction in the neuroscience community, Deisseroth is taking the technology a step further. He sees patients at Stanford Medical School and, using optogenetic approaches, his lab establishes different animal models of neuropsychiatric disease.

"We now have models of Parkinson's disease, depression and altered social behaviour relevant to autism," notes Deisseroth. In all of these disorders, the circuit dynamics are not working well, so Deisseroth's goal is to use his optogenetic models to deconstruct the neural circuitry in an attempt to better understand which parts of the brain are functioning properly, and which are not, in the different disease states.



K. DEISSEROTH

Light is used to open ion channels and activate specific neurons in the brain.

With the ability to work on a wide range of organisms, an expanding tool kit, and growing interest among researchers, Deisseroth sees even greater possibilities for the system. "In the end, if you can get the gene in, along with the proper light, it works very well." N.B.

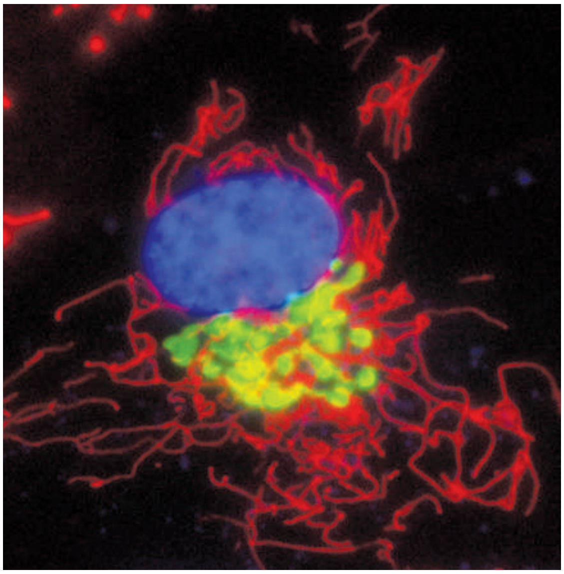
fluorescence spot that, when scanned across the specimen, gives super-resolution images automatically.

"STED was the first concept that really showed you could go beyond the barrier of half the wavelength of light in the focal plane," says Hell. More and more super-resolution approaches are starting to be described in scientific journals as researchers see the potential of the method.

For detecting single molecules in cells, William Moerner, a chemist at Stanford University, says photoswitching of single molecules has created a real sense of excitement in the community. By exploiting the ability of fluorophores to switch states, researchers can work with high concentrations of labelling products — crucial for imaging of structures. And, by having most of the labelled molecules 'off' and only a few 'on' at a given moment, they can improve the resolution down to that single-molecule level of detection.

This premise is at the heart of several newer super-resolution approaches. These include photoactivated localization microscopy[1] (PALM) developed by Eric Betzig at the Howard Hughes Medical Institute's Janelia Farm Research Campus in Ashburn, Virginia, and his colleagues; fluorescence-PALM, developed by Samuel Hess and his colleagues at the University of Maine in Orono[2]; and stochastic optical reconstruction microscopy[3] (STORM), developed by Xiaowei Zhuang and her colleagues at Harvard University. These methods can now image structures in cells at a resolution far below 200 nanometres.

Even though fluorophores are used in all these methods, both new and old photoswitchable molecules are proving useful for super-resolution imaging. Although Moerner and his colleagues recently used the readily available enhanced yellow fluorescent protein (EYFP) to image proteins at a resolution of less than 40 nanometres in *Caulobacter crescentus*[4], he is also involved with the National Institutes of Health Molecular Imaging Exploratory Center programme in which his group, along with others at Stanford University and Kent State University in Kent, Ohio, are developing new dyes and fluorophores specifically for single-molecule imaging. The centre's efforts are starting to pay off. Researchers there recently invented a new class of photoswitchable molecules based on the switching of an azide functional group to an amine when exposed to activating light beams[5], as well as covalently linking cyanine dyes to a molecule to improve switching performance in STORM experiments. "Although I know that many groups are working to optimize fluorescent proteins for switching, now the community can work on optimizing small molecules for switching as well," says Moerner.



INVITROGEN

Organelle-specific probes can illuminate cellular structures.

Higher resolution for all

For those researchers interested in breaking the diffraction barrier in their own research, the road to super resolution will become easier in the coming years. "Initially, I think there was quite a reluctance to believing that you would get super resolution," says Hell. When he started developing STED microscopy, the initial system relied on a series of flexible lasers, creating a complex system that many outside researchers initially thought would be required for any STED imaging application. But today, Hell says the development of laser systems that are easier to use and have the right light structure to perform STED, alongside ever-improving confocal systems with their spectral-imaging capabilities for those interested in the PALM and STORM approaches, are encouraging researchers to try super-resolution imaging.

Leica Microsystems is the only company to offer STED capability for super-resolution imaging. The Leica TCS STED Superresolution Microscope is a broad-band confocal platform with multi-photon capability that integrates the STED concept, allowing users to obtain resolutions of better than 100 nanometre. Hoppe says that it took four years for the company to develop a STED system that was both stable and user-friendly.

"For STED to work, the two beams — for excitation and depletion — need to be perfectly aligned in space and time," says Hoppe. This is crucial to super resolution with STED — if the lasers are out of alignment, researchers cannot obtain super-resolution imaging.

To deal with this, engineers at Leica developed fully automated beam-correction functionality on the TCS STED system. If an environmental factor — for example, the temperature of the room — changes, researchers can simply push a button and the system realigns the two lasers automatically.

Advances in fluorescent probes and microscope instrumentation are pushing resolution barriers and multicolour imaging possibilities. So this year, as microscopists around the globe submit the amazing pictures they capture of the cellular world to those annual imaging contests — to the delight of scientists and the public alike — many of those same entries could produce fundamental insights into the inner workings of the cell at a level of resolution that previously had not been possible. The future is bright. ■

Nathan Blow is the technology editor for *Nature* and *Nature Methods*.

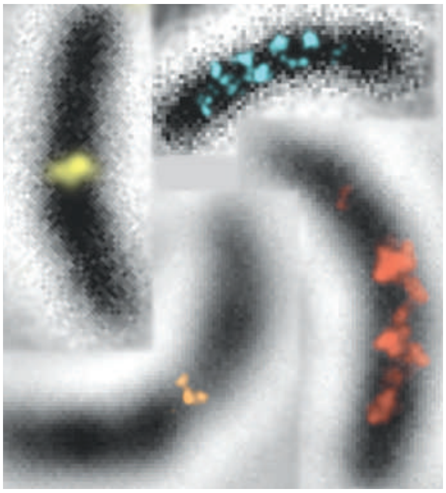
LEICA MICROSYSTEMS

Leica Microsystems recently developed a commercial version of Stefan Hell's STED platform.

1. Betzig, E. *et al. Science* **313,** 1642–1645 (2006).
2. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. *Biophys. J.* **91,** 4258–4272 (2006).
3. Rust, M. J., Bates, M. & Zhuang, X. *Nature Methods* **3,** 793–795 (2006).
4. Biteen, J. S. *et al. Nature Methods* **5,** 947–949 (2008).
5. Lord, S. J. *et al. J. Am. Chem. Soc.* **130,** 9204–9205 (2008).
6. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. *Nature Neurosci.* **8,** 1263–1268 (2005).

NATURE|Vol 456|11 December 2008

J. BITEEN



Fluorescence switching is key to single-molecule imaging methods such as PALM and STORM.

to advance the use of FRET in high-throughput screening applications. Here, the challenge of obtaining a strong FRET signal is crucial to overcoming any background fluorescence. Because the simple approach of making a brighter fluorescent protein can be time-consuming, another approach, being used by a number of developers, is to take advantage of the 'states' — the differences between the excitation and emission — of a fluorescent protein. Time resolved FRET (TR-FRET) uses long-lived fluorophores that exhibit a time delay between their excitation and emission, minimizing potential interference from background fluorescence. Several companies now offer TR-FRET platforms for cellular interaction assays, including Life Technologies, Cisbio in Bedford, Massachusetts, Covalys of Witterswill, Switzerland, and PerkinElmer located in Waltham, Massachusetts.

Quantum of happiness

Another route to fluorescence is provided by quantum dots, which "present better photostability than organic dyes and allow for increased multiplexing capability", says Stephen Chamberlain, manager of business development at Life Technologies.

Quantum dots are composed of a semiconductor core shell surrounded by an exterior coating. What makes them unique as fluorophores is their ability to be excited by a single wavelength of light — the emission from a quantum dot is dictated by the size of its core shell. As the size of the core shell can be varied to obtain different light-emission spectra, quantum dots present a particularly interesting solution for researchers interested in experiments using multiple coloured fluorophores.

But unlike fluorescent proteins, quantum dots are not genetically encodable, which has been a stumbling block to more widespread usage. "It is hard to get them into the cell," says Alice Ting, a chemist at the Massachusetts Institute of Technology in Cambridge, who has been working with quantum dots for several years. "And then once inside, how do you target them to a specific protein?"

Most researchers and developers are looking towards the outer coating for targeting. At Life Technologies, says Chamberlain, they place a polymer coating around the core shell of their Qdot Nanocrystals along with an additional layer of polyethylene glycol on top of the polymer coating to reduce the sticking together of the Qdots, and to provide a surface for the attachment of antibodies and other molecules for targeting.

Despite the challenges of targeting, quantum dots are now being tried in a number of imaging applications. Chamberlain says that like their organic-dye counterparts 20 years ago, researchers are still in the process of developing robust application protocols for quantum dots. Although quantum dots can be used for FRET applications, much like fluorescent proteins, they tend to make better donors in a FRET interaction because they can be excited by light that is blue-shifted relative to their emission. They are being used for *in vivo* imaging applications, such as looking at blood vasculature, because of their high residence times and ability to excite in the near-infrared region of the spectrum.

Ting says using quantum dots for single-molecule imaging is a much more straightforward application at the moment, although the issues of delivery and targeting still exist, and several groups are now taking advantage of the greater brightness of quantum dots to track the movement of single molecules within cells.

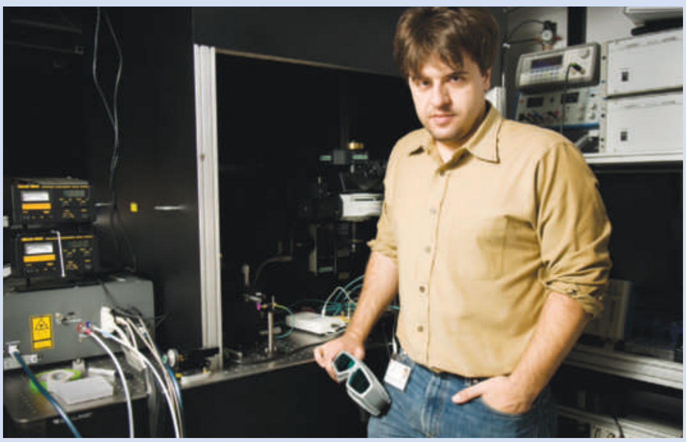
Microscopes respond

As the types and range of applications of fluorescent probes and quantum dots continue to grow, developers of confocal microscopes have been working hard to keep instrumentation

LIGHT ACTIVATED

When scientists first discovered light-absorbing molecules — in microorganisms — that could control the flow of ions into a cell, many researchers became interested in comparing how these proteins differed from their visual pigment counterparts. But Karl Deisseroth at Stanford University in California saw something else in these light-activated channels — a potential tool to dissect the function of the brain.

Deisseroth — along with Edward Boyden, his postdoctoral fellow at the time, now at the Massachusetts Institute of Technology in Cambridge, and his graduate student Feng Zhang — collaborated with microbiologists to hunt for these light-activated channels from various species. "I come from a synaptic physiology background, so we just sequentially tried these and other proteins to find the ones that worked the best," says Deisseroth. It was a collaboration with Georg Nagel and Ernst Bamberg at the Max Planck Institute of Biophysics in Frankfurt, Germany, who supplied a clone of a gene called channelrhodopsin-2 from *Chlamydomonas reinhardtii* to Deisseroth that opened the floodgates.



K. DEISSEROTH

Karl Deisseroth develops optogenetic methods for controlling neural activation.

Channelrhodopsin-2 is a gated light-sensitive cation channel that uses a molecule of all-*trans* retinal to absorb photons. When Deisseroth, Boyden and Zhang expressed channelrhodopsin-2 in hippocampal neurons in the mouse brain then shone blue light on the region, they found the cells with channelrhodopsin-2 responded to the light stimulation, opening the channel and initiating the flow of ions, which resulted in an action potential in those neurons[6]. Deisseroth says that one of the most pleasant surprises to emerge from that first series of experiments turned out to be the precision of the system. "Even though these molecules are not designed to generate action potentials and work on this time scale in neurons," he says, "it turns out that they can."

This result led to the start of a new field, coined 'optogenetics' by Deisseroth in 2006, where researchers are combining optics with genetics to explore the workings of neural circuits. In the years since the first description of channelrhodopsin-2, Deisseroth's lab has gone on to advance the system and develop new probes. The team identified and developed two more light-activated

up to the challenge. "In general, researchers are still requesting the same three things: higher speed, better signal-to-noise and lower toxicity," says Michael Davis, a confocal product manager from Nikon Instruments in Melville, New York. But he is quick to add that there are several emerging fluorescent-protein techniques that have required more sophisticated confocal technology.

Davis says more researchers are now using photoactivatable and photoconvertible proteins in their research. The idea here is to use a specific wavelength of light to excite the fluorescent protein, which can then be turned off, in the case of photoactivatable proteins, or even converted to another colour, in the case of photoconvertible proteins, with a different wavelength of light. Evrogen now offers both kindling red fluorescent protein (KFP-Red), which can be either reversibly or irreversibly activated for either short or long-term protein tracking in a cell, and PS-CFP2, which is a photoconvertible protein that will switch from cyan to green in response to light activation. And MBL International in Woburn, Massachusetts, offers the photoconvertible proteins CoralHue Kaede and CoralHue Kikume, which convert from green to red in response to light activation.

The development of these fluorescent proteins has led to dual scanning systems for confocal microscopes, allowing for the simultaneous imaging and activation of photoactivatable and photoconvertible proteins, along with new spectral imaging technologies that can accurately assay a broad range of possible spectral emissions.

Another trend in the development of confocal instrumentation is the emergence of resonance scanners. "People would like to see the resolution they get out of a point scanner, but with the speed of a field scanner," says Davis. Several developers hope that resonance scanning will bridge this divide. With a resonance-scanning confocal system, the scanner mirrors are moved at extremely high frequencies, permitting much higher frame rates and minimizing the time the scanner dwells at any particular point in the sample, allowing researchers to study more dynamic events in a cell while reducing the effects of phototoxicity on it. Both Nikon and Leica Microsystems have introduced resonance-scanning confocal systems in recent years.

S. HELL

Stefan Hell developed the STED imaging approach.

Small switches

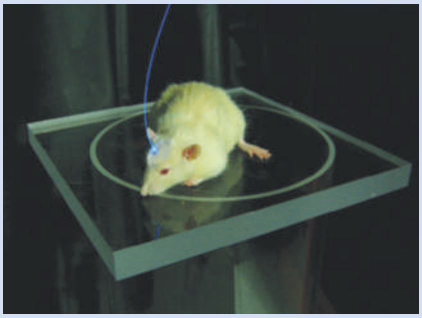
Although dual imaging and faster scanning are enhancing fluorescence imaging, the next step in improving the optical resolution of microscopes might require all developers to look more closely at the properties of those fluorescent proteins. "Optics alone can only provide resolution to a certain limit," says Davis.

The ability of a fluorescent protein or a dye to switch states is also at the heart of an emerging field in microscopy. "Fluorescence switching is the common basis to all super-resolution methods that have been devised so far," says Stefan Hell, director of the Max Planck Institute for Biophysical Chemistry in Gottingen, Germany, who created a super-resolution imaging method known as stimulated emission depletion (STED) 'nanoscopy'.

Before super-resolution imaging, the resolution limit of optical microscopy was thought to be determined by the diffraction barrier of light. For example, in a high-resolution confocal microscope, a beam of light is focused down to a spot on the focal plane of about 200 nanometres in size. Because all molecules within that spot are excited in parallel, the spatial resolution is limited to about this value. Hell's approach with STED was to use the transition states of fluorophores, or their photoswitching capabilities, to effectively reduce the size of that spot — potentially all the way down to a single molecule. In a STED microscope, the focal spot is overlapped with a ring of light that switches off fluorescence everywhere in the spot except at the centre of the ring. The outcome is a much smaller

molecules for optogenetic control of neurons. These were an inhibitor called halorhodopsin, and a *Volvox* channelrhodopsin that can also initiate neuronal activity, but is more than 100 nanometres red-shifted from the peak of the *Chlamydomonas* channelrhodopsin, allowing separable channels of optical control. They also developed various targeting strategies to get channels expressed in specific neuronal cell types and populations, and fibreoptic/laser-diode hardware for adaptation to mammals.

The growing number of researchers interested in applying optogenetic approaches to their particular research questions has led to an unexpected effort within Deisseroth's lab. "Technology distribution and helping other researchers has been another big part of what we are doing," says Deisseroth. His lab has supplied clones of the channels to more than 300 other labs around the globe for use in organisms ranging from mice to fruitflies. Although the channels have worked in every species tried thus far, in some cases, minor modifications have been required. Although there is sufficient all-*trans* retinal in the brains of mice for the channels to function properly, invertebrates must be supplemented with all-*trans* retinal through their diet to achieve channel activation.

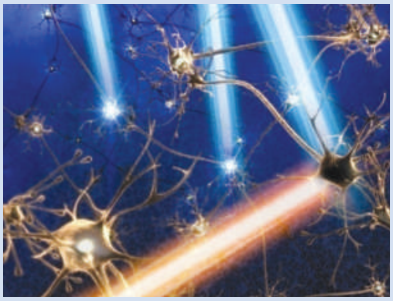


K. DEISSEROTH

Mouse brain neurons can be stimulated by light from fibre-optic cables attached to the animal.

As the basic use of these light-activated channels for studying brain function and circuitry gains more traction in the neuroscience community, Deisseroth is taking the technology a step further. He sees patients at Stanford Medical School and, using optogenetic approaches, his lab establishes different animal models of neuropsychiatric disease.

"We now have models of Parkinson's disease, depression and altered social behaviour relevant to autism," notes Deisseroth. In all of these disorders, the circuit dynamics are not working well, so Deisseroth's goal is to use his optogenetic models to deconstruct the neural circuitry in an attempt to better understand which parts of the brain are functioning properly, and which are not, in the different disease states.



K. DEISSEROTH

Light is used to open ion channels and activate specific neurons in the brain.

With the ability to work on a wide range of organisms, an expanding tool kit, and growing interest among researchers, Deisseroth sees even greater possibilities for the system. "In the end, if you can get the gene in, along with the proper light, it works very well." N.B.

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Fluorescent probes and quantum dots			
AnaSpec	Fluorescent dyes and physiological indicators; FRET probes; labelling kits for proteins and peptides	San Jose, California	www.anaspec.com
Biotium	Fluorescent dyes and indicators; cell-biology and apoptosis assay kits	Hayward, California	www.biotium.com
BioVision	Fluorescent protein vectors, antibodies, tools and kits for cell-biology research	Mountain View, California	www.biovision.com
Cisbio	Time-resolved FRET assays	Bedford, Massachusetts	www.htrf.com
Clontech	Fluorescent proteins and reagents, fluorescent reporter systems	Mountain Valley, California	www.clontech.com
Covalys	Self-labelling fluorescent tags, TR-FRET systems, protein interactions systems	Witterswill, Switzerland	www.covalys.com
Denovo Biolabels	Development of fluorescent markers and dyes for bioanalytical applications	Münster, Germany	www.biolabels.com
eBioscience	eFluor high-yield quantum nanocrystals for imaging applications	San Diego, California	www.ebioscience.com
Evrogen	Fluorescent proteins and molecular-biology systems	Moscow, Russia	www.evrogen.com
GE Healthcare	Fluorescent proteins for cell analysis and drug-screening applications	Piscataway, New Jersey	www4.gelifesciences.com
Invitrogen	Fluorescent proteins, dyes and quantum dots	Carlsbad, California	www.invitrogen.com
LI-COR Biosciences	Near-infrared dyes for labelling antibodies and other proteins; *in vivo* imaging applications using near-infrared dyes	Lincoln, Nebraska	www.licor.com
LUX biotechnology	GFP vectors and proteins, phosphoproteins and luciferase vectors and proteins	Edinburgh, UK	www.luxbiotech.com
MBL International	Fluorescent probes and photoconvertable proteins	Woburn, Massachusetts	www.mblintl.com
Mo Bi Tec	Fluorescent dyes and conjugates; fluorescent calcium indicators	Göttingen, Germany	www.mobitec.de
Promega	HaloTag system for protein tagging	Madison, Wisconsin	www.promega.com
VisEn Medical	Probes to detect matrix metalloprotease and cathepsin K activity; tomography imaging systems	Bedford, Massachusetts	www.visenmedical.com
Cell imaging and microscopy			
Applied Precision	Restoration microscopy, high-content cell imaging and image-analysis tools	Issaquah, Washington	www.api.com
Carl Zeiss	Microscopes, digital imaging of cells, image-analysis software	Jena, Germany	www.zeiss.com
Chroma Technology	Optical filters	Rockingham, Vermont	www.chroma.com
Cytonome	High-speed optical cell sorting	Boston, Massachusetts	www.cytonome.com
Dako	Automated cellular imaging, fluorescent-activated cell-sorting instruments and kits	Carpinteria, California	www.dakousa.com
Fujifilm	Imaging systems for array analysis	Stamford, Connecticut	www.fujimed.com
Hamamatsu	Imaging systems	Hamamatsu City, Japan	www.hamamatsu.com
Improvision	Three-dimensional imaging software	Coventry, UK	www.improvision.com
Intracellular Imaging	Digital fluorescence imaging and photometry systems	Cincinnati, Ohio	www.intracellular.com
ISS	Spectrofluorometers for applications including time-resolved confocal microscopy, life-time imaging and fluorescence-correlation spectroscopy	Champaign, Illinois	www.iss.com
Leica Microsystems	Light microscopes, stereomicroscopes and confocal microscopes; imaging-analysis software	Wetzlar, Germany	www.leica-microsystems.com
Lightools Research	*In vivo* imaging tools, systems and filters	Encinitas, California	www.lightools.com
MAG Biosystems	Microimaging systems for FRAP; multichannel-imaging and spectral-imaging systems	Pleasanton, California	www.magbiosystems.com
Mauna Kea Technologies	Miniaturized fluorescence confocal microscopy system for *in vivo* intra-tissue imaging	Paris, France	www.maunakeatech.com
Micro Video Instruments	Microscopes and accessories	Avon, Massachusetts	www.mvi-inc.com
MMI Molecular Machines and Industries	Equipment for micromanipulation and microdissection	Glattbrugg, Switzerland	www.molecular-machines.com
Nanoptek	Digital tunnelling microscopes	Maynard, Massachusetts	www.nanoptek.com
Ningbo Kingstic	Stereomicroscope, digital microscopes and epi-fluorescence microscopes and accessories	Ningbo, China	www.biomicroscope.com
Nikon	Imaging solutions, microscopes, objectives, digital cameras, imaging software and analysis tools	Melville, New York	www.nikoninstruments.com
Olympus America	Microscopes and imaging systems; digital imaging; *in vivo* imaging solutions	Center Valley, Pennsylvania	www.olympusamerica.com
Omega Optical	Optical filters and coatings for imaging; custom filters	Brattleboro, Vermont	www.omegafilters.com
Optronics	Digital cameras for microscopes; image-analysis software	Goleta, California	www.optronics.com

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL	
Princeton Instruments	Equipment for imaging and spectroscopy	Trenton, New Jersey	www.princetoninstruments.com	
Semrock	Optical filters for microscopy applications	Rochester, New York	www.semrock.com	
UVP	Instruments for *in vivo* whole-animal imaging using fluorescent proteins	Upland, California	www.uvp.com	●
Visiopharm	Software packages for the analysis and management of imaging data	Hørsholm, Denmark	www.visiopharm.com	

General

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL	
Applied BioPhysics	Automated instruments for cell monitoring and electric cell-substrate impedance sensing	Troy, New York	www.biophysics.com	
ALEXIS Biochemicals	Suppliers of reagents for molecular and cell-biology research	Lausen, Switzerland	www.alexis-corp.com	
AMS Biotechnology	Antibodies to GFP, antibody microarrays, tools for cell-signalling research and protein biochemistry	Abingdon, UK	www.amsbio.com	
Beckman Coulter	Automated tools for genomics and proteomics research	Fullerton, California	www.beckmancoulter.com	
BioFlow Technology	Developing bioreactor systems for the culturing of cells in a three-dimensional environment	Novi, Michigan	www.bioflowtech.com	
Biomol	Services for chemical synthesis, cell culture and antibody production	Hamburg, Germany	www.biomol.de	
Bio-Rad	Products, instruments and software for life-sciences research	Hercules, California	www.bio-rad.com	
BMG LABTECH	Microplate and array readers and handling systems	Offenburg, Germany	www.bmglabtech.com	
Brinkmann Instruments	Laboratory instrument supplier; consumables	Westbury, New York	www.brinkmann.com	
Cambrex	Products for molecular and cell-biology research	East Rutherford, New Jersey	www.cambrex.com	
EMD Biosciences	Calbiochem, Novabiochem and Novagen product lines	San Diego, California	www.emdbiosciences.com	
Eppendorf	Consumables for molecular biology; instrumentation	Hamburg, Germany	www.eppendorf.com	
EUGENEX Biotechnologies	Development of test cell lines	Tägerwilen, Switzerland	www.eugenex.com	
Gilson	Pipettes, automated liquid handling, liquid-chromatography systems and software	Middleton, Wisconsin	www.gilson.com	
Hamilton	Automated liquid-handling stations	Reno, Nevada	www.hamiltoncompany.com	
Harvard Apparatus	Instruments and equipment for electrophysiology and cell biology	Holliston, Massachusetts	www.harvardapparatus.com	
Horiba Jobin Yvon	Spectroscopy systems and accessories including Raman, atomic-emission and UV spectroscopy	Edison, New Jersey	www.jobinyvon.com	
Integra Biosciences	Equipment for sterilization, liquid handling, cell culture and sample storage	Chur, Switzerland	www.integra-biosciences.com	
Irvine Scientific	Defined media for cell-culture applications; custom media services	Santa Ana, California	www.irvinesci.com	
Merck	Chemicals, kits and reagents for molecular- and cell-biology-related research	Darmstadt, Germany	www.merck.de	
Molecular Devices	Liquid-handling and microplate-processing equipment; imaging instruments	Sunnyvale, California	www.moleculardevices.com	
MP Biomedicals	Reagents and chemicals for research	Santa Ana, California	www.mpbio.com	
Nalge Nunc International	Labware	Rochester, New York	www.nalgenunc.com	
New England Biolabs	Molecular-biology-related reagents, kits and enzymes	Ipswich, Massachusetts	www.neb.com	●
PerkinElmer	Instruments, reagents and kits for life-sciences research	Waltham, Massachusetts	las.perkinelmer.com	
Roche Diagnostics	Reagents and kits for molecular biology; genomics instrumentation and software	Burgess Hill, UK	www.roche-applied-science.com	●
Stratagene	Tools and reagents for molecular-biology, genomics and proteomics research	La Jolla, California	www.stratagene.com	
Takara Bio	Reagents, kits and consumables for molecular biology	Shiga, Japan	www.takara-bio.com	●
Thermo Fisher Scientific	Chemicals and reagents for life-sciences research; instruments, robotics and automation	Waltham, Massachusetts	www.thermofisher.com	
Tocris Bioscience	Chemicals for life-science research; contract research services	Bristol, UK	www.tocris.com	
Wako Chemicals USA	Speciality chemicals; clinical diagnostic reagents	Richmond, Virginia	www.wakousa.com	
USB	Chemicals and reagents for molecular biology	Cleveland, Ohio	www.usbweb.com	

● see advertisement

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

One stereotypical public perception of scientists is of stoic, noble figures who don lab coats and work religiously for years on end until they are rewarded with a major discovery. This image is, of course, flawed for many reasons — one being that scientists, like everyone else, are emotional about their work. The lucky ones are passionate about it. And emotion is an important force motivating experiments and career decisions.

This week we delve into the fossil-fuels industry and explore opportunities available to geologists, chemists, physicists and computer scientists (see page 832). For many, such careers can be rewarding, financially and otherwise. Much of this work has a satisfying problem-solving element. The aim is clear, as are the rewards: find the oil (or a better way to extract it) and your company will profit — potentially in a big way. And with the demand for new recruits still high despite the economic downturn, salaries should stay sizeable.

Some scientists, though, might have reservations about such work. Oil and coal companies can be seriously detrimental to the environment and to wildlife. Those who feel science is a means to work towards a greater good — curing disease or solving the energy crisis — might frown on such vocations.

Krystal Walker, a student at the University of Utah's mining programme in Salt Lake City and a self-identified environmentalist, decided she was going to try to change these companies' practices from the inside, by finding ways to minimize their impact on the landscape. This, of course, is a bold gambit with no guarantees.

Those with a passion for saving our planet or benefiting society may be better served by a job in a non-governmental organization (see *Nature* **455,** 1002–1003; 2008). As Anumita Roychowdhury, associate director of the Centre for Science and Environment (CSE) in Delhi, India, put it in our recent story: the CSE looks for people who are a "little mad about things". For those people, the practical problem-solving of the oil and coal industries might not be quite enough.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Assistant European Manager:
Nils Moeller (4953)
***Nature*events:**
Ghizlaine Ababou (+44 (0) 20 7014 4015)
e-mail: g.ababou@nature.com
Southwest UK/RoW:
Alexander Ranken (4944)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
North Germany/The Netherlands/Eastern Europe: Kerstin Vincze (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

OLD FUELS, NEW WEALTH

Staff shortages in the US fossil-fuel industry are good news for geologists, chemists and physicists. The boom is likely to continue despite the economic downturn and the rise of renewable energy, says **Emma Marris**.

AP PHOTO/J. MCINTOSH

As Garrett Vice approached the end of his master's degree at the University of Nevada Mackay School of Earth Sciences and Engineering in Reno, the job offers started arriving. He found potential opportunities at geothermal companies, mining companies, oil companies and state agencies. Vice settled on a position with ExxonMobil in Houston, Texas, won over by on-the-job training that combined mentoring and two years of classes in geophysics. "It is phenomenal," he says of the training, which focuses on practical problem-solving. "Every day I feel like Sherlock Holmes." And the salary, says Vice, was higher than he expected. He and his wife have already bought a house in Houston.

The fossil-fuel industry is hungry for more people like Vice. Pay for geologists in the US petroleum industry has risen by well over a third since 2005, according to a survey by the American Association of Petroleum Geologists (see http://tinyurl.com/6mxfry). New hires, typically with master's degrees, can expect to start at about US$82,000 a year plus a signing bonus. In comparison, the American Geological Institute reports that environmental or government jobs for geoscience master's graduates pay about $45,000 a year (see http://tinyurl.com/66b88w).

"It is one of the fulfilling jobs a scientist can have. You are actually doing what you went to school for."
— Adam Seitchik

Purveyors of oil, gas and coal also need chemists, physicists and computer scientists. Generally, the geoscientists and physicists work on 'upstream' jobs: finding and extracting supplies. Chemists often work on 'downstream' jobs, including dividing crude oil into multiple streams and transforming each into fuels and other commercial products. Computer scientists work on software to improve and accelerate both downstream and upstream tasks.

But the highest demand for expertise is in the geosciences. And most industry insiders insist that the boom will continue for the foreseeable future. Mike Nelson, chairman of the mining engineering department at the University of Utah in Salt Lake City, says that at least one mining company wooed his students with starting bonuses totalling $10,000 a year for the first five years. "They were so desperate to get people to work in their mine in Wyoming," he says.

Chemists and chemical engineers are in high demand as well, according to Bill Lemons, regional director for staffing firm Kelly Scientific Resources in Houston, Texas. They typically help improve the manufacture of fuels or develop new polymers and membranes for different commercial products. The American Chemical Society's 2006 salary data show chemical engineering PhDs fetching a median starting salary of approximately $88,000 in the oil industry.

Making way for the next boom

The reason for this vigorous demand? The large number of scientists trained in the 1960s and 1970s are reaching retirement age. And lay offs in the subsequent two decades have left few mid-career workers ready to replace them. This has created "a gap of competence", according to James Taranik, director of the Mackay School. "In the science disciplines, we are starting to see a demand that is far outstripping supply," says Taranik. "It is changing practically instantaneously as the baby-boom generation is beginning to retire."

High energy prices — peaking in July with oil at $148 a barrel — have helped pay for new salaries. But September's economic upheaval and sub-$50-a-barrel oil have yet to change hiring trends. Taranik suggests that, although the economic slowdown may be causing a slight sag in the industry, strong growth in China, India and Brazil will offset this in the long-run. Oil companies are often very conservative when making business decisions, ignoring short-term fluctuations in oil price and assuming a conservatively low per-barrel base price. "Most companies never used a price much over $40 in calculating their economics," says executive recruiter Mike Ayling of MLA Resources in Tulsa,

Oklahoma. He adds that most companies have been "operating so thin on expertise that there is little room to lay everyone off".

The worldwide operations of BP, for example, hire about 650 college graduates a year. BP spokesman Robert Wine estimates that two-thirds or more are scientists. And Wine is confident that trend will continue. "These numbers are not affected by the current climate," he says. "Our projects are 20-, 30-, 40-year investments, so we need to keep the recruitment of the next generation well up." A smaller oil company, such as Devon Energy of Oklahoma City, has about 300 geoscientists on the payroll at any time.

There are also 'service companies' such as Halliburton, Baker Hughes and Schlumberger, all based in Houston, which provide specific skills and data to the major companies on a contract basis, and hire numerous chemists, geoscientists, computer scientists and physicists. For example, Schlumberger, which spent $728 million on research and development (R&D) last year, collects and sells seismic profiles of different regions, provides companies with real-time data streams about new holes while they drill, manages projects and sells specialized tools.

An insatiable thirst for energy has also meant new frontiers and new opportunities. Opened in the late 1970s, Alberta's oil sands in Canada (pictured above) are bustling with activity. The recoverable bitumen, a type of heavy petroleum mixed with dirt, represents 173 billion barrels of oil. That's the most oil in one place outside Saudi Arabia, according to Alberta's government (see http://tinyurl.com/6d4x32). The largest oil-sands company, Syncrude, based in Fort McMurray, spends Can$40 million (US$32 million) a year on R&D, has about 100 scientists on staff, and funds research at institutions such as the University of Alberta in Edmonton and the University of Waterloo in Ontario. Companies want to become more efficient in recovering bitumen, a process that took a leap forwards in the mid-1990s, when Syncrude researchers invented a way to pipe oil sands mixed with water over long distances for processing. Syncrude has been tweaking this system ever since. The company also needs chemists to work on ways of limiting wear and tear on equipment; oil sands corrode the teeth on the enormous shovels.

Practical problem solving attracts many scientists to oil exploration. "It is one of the fulfilling jobs a scientist can have," says Adam Seitchik of Devon Energy. "You are actually doing what you went to school for in your day-to-day job." Until recently switching to a managerial role, Seitchik worked with data provided by service companies: seismic images of the soil beneath kilometres of water in the Gulf of Mexico. Aided by special software, Seitchik would interpret the processed data and generate maps of geologic structures that might hold or trap hydrocarbons migrating upwards in porous rock.

M. NELSON

University of Utah students go underground.

"Our students can expect to work on most of the continents of the world. The adventurers just can't wait to go."
— James Taranik

Digging for coal

Exploration for coal is generally more straightforward than for oil, as coal is often found near the surface over a large area — sometimes even exposed on a hillside. But mining it can be just as tricky. Nevertheless, many jobs entail what Nelson describes as "day-to-day getting the rock out of the hole". Geologist Terry Rowland and his crew, who work for the Coteau Properties Company in North Dakota, routinely drive out on the rolling prairie ahead of ongoing mining operations. They bore holes in the ground and lower probes in. The measurements they get back — including electrical resistance, density and natural radioactivity — are used to advise the drivers of heavy machinery. "It is not bad, except for when the wind blows, which is almost all the time," says Rowland, who entered the field in 1976 with a bachelor's degree in geology. At larger offices, he adds, "the younger ones go out, and the older guys stay in the office. But I am kind of a lone ranger out here."

Rowland is a survivor. He has made it through three rounds of lay offs — once when 17 out of 20 colleagues lost their jobs. Despite bold projections about continued demand for scientists, the energy business is cyclical, as the lean years in the 1980s and 1990s demonstrated. "There will be down years and there will be boom years," admits Lemons, echoed by many other industry insiders.

Rowland says researchers must be willing to roll with the changes and the inconveniences. In 1990, his company moved him from its Bismarck offices to the mine near the hamlet of Beulah, 86 miles away. He arranged a four-day week and rented a small apartment in Beulah, to limit commuting. On Thursdays, he heads home for three days with his family in Bismarck.

Many in the industry will be asked to make bigger moves. According to Lemons, energy-industry scientists are an itinerant band, with Houston as a crossroads for many. "That is absolutely part of the package," Lemons says. "They'll end up in Asia, the Middle East or Africa, and we have a lot of people that travel back and forth to Alaska."

Taranik says "I tell our students that they can expect

A Devon Energy/ConocoPhillips oil rig in the Gulf of Mexico.

DEVON ENERGY

to work on most of the continents of the world. But we do find out that there are students who don't want to work overseas because they are afraid of terrorists or malaria or dengue or hepatitis C. Some of the students are reluctant to be adventurers. The adventurers just can't wait to go to these places."

But timidity isn't the only reason to think twice about jobs in the industry. The environmentally minded could have qualms about working for companies that profit from greenhouse-gas-emitting energy sources — companies with a reputation for polluting and also for taking advantage of small communities.

Krystal Walker has "a green streak", but learned of the high demand for mining engineers a couple of years back while working, unhappily, as an administrative assistant after she had dropped out of a master's in biology. "I was horrified by the idea — 'I can't go into mining! That's going to the dark side,'" says Walker. Eventually, though, she decided it would be "easier to make changes from the inside than from the outside". She's now in her third semester at the University of Utah's renowned mining programme and hopes to use her biology training to help companies minimize damage to the landscape and facilitate reclamation of mined sites. Her husband is supporting her while she studies, but she expects the financial sacrifice to be worth it: she made a deal with her husband. I could do the course "if I would make in two years more than I would have made in the three years I was in school," she says.

Keith Slack, manager of the extractive industries programme at Oxfam America, is constantly on the lookout for people with technical training who want to work with environmental or community groups. In the case of a dispute with an energy company involving its refinery or mining site, they might help the locals set up water-quality monitoring or understand technical documents such as environmental-impact statements. Some of these jobs are volunteer positions for non-profit organizations supporting community groups.

OXFAM AMERICA

'It may not be easy to find classes on social and environmental issues. You may have to register at other schools.'
— Keith Slack

But there are paid positions for technical experts such as consulting geologists at outfits such as the Center for Science in Public Participation in Bozeman, Montana, or Engineers Without Borders in Cambridge, UK.

Slack helps inform students at mining schools about some of the environmental and human-rights aspects of their future careers. He suggests that students educate themselves about what he sees as "increasingly bottom-line issues": more and more, international laws have made behaving as an irresponsible corporate citizen bad for business.

"It may not be easy to find classes to give them training on social and environmental issues," Slack says. "They may have to cross-register at other schools or branch out onto other departments."

Stepping between old and new

Some of the more environmentally conscious might choose to pursue a career path in the renewable-energy industries, which are likely to see continued growth. But although renewable energy has been the subject of much political wrangling and media attention, jobs related to traditional fuels are likely to offer promising career paths for many years to come.

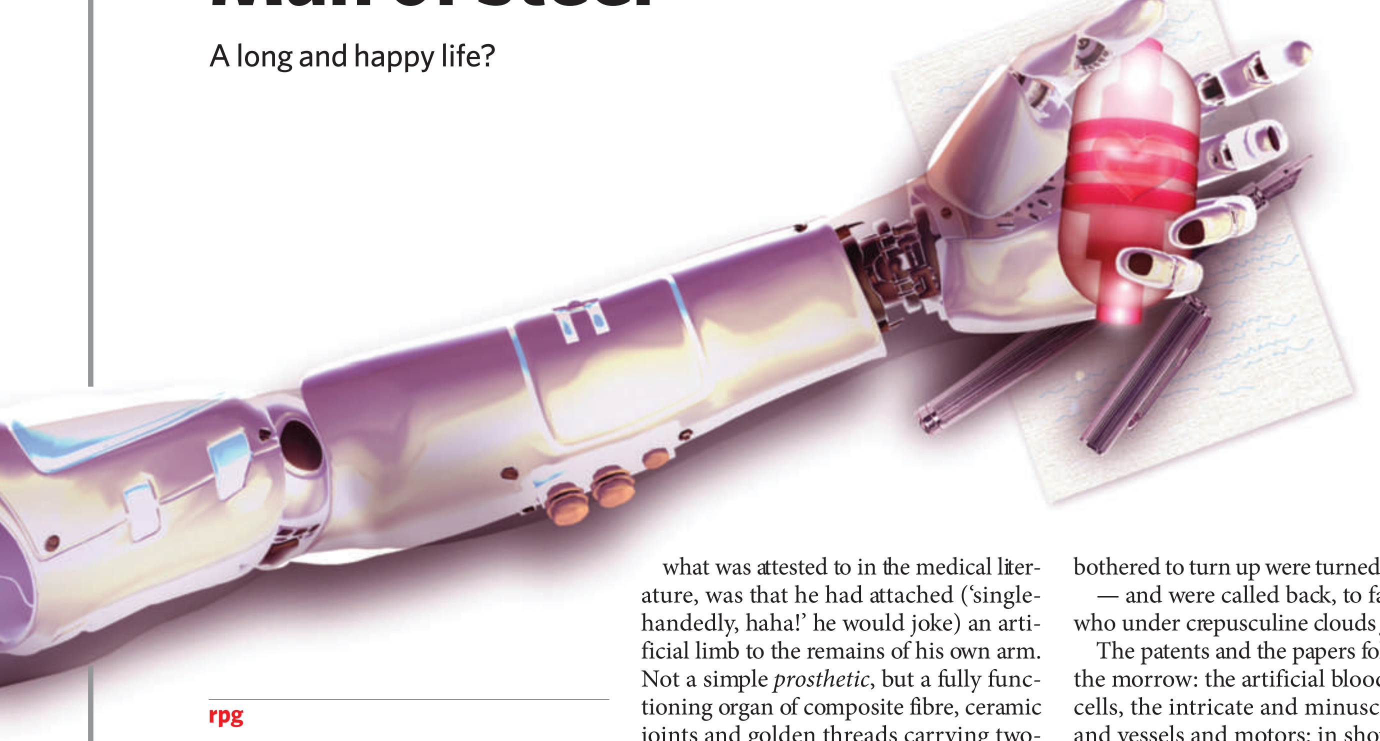
Fossil-fuel industry jobs may use skills that are transferable to renewables. An oil-industry geologist could turn to finding promising sources of geothermal energy; a chemist could work to make reactions more efficient at a biofuels company. A scientist could conceivably bounce back and forth between the old industries and the snazzy green newcomers, even within the same large, energy company.

Vice, currently gaining geophysics expertise from his near-retirement mentors at ExxonMobil, originally hoped for a job in geothermal energy. Perhaps in ten years, as the push for alternative energies continues, he will use that expertise to hop from one energy stream to another, as he pleases. ■

Emma Marris is a correspondent for *Nature* based in Columbia, Missouri.

Man of steel

A long and happy life?



JACEY

rpg

The suicide note itself wasn't particularly remarkable.

Handwritten, of course. Even the oldest computers would have detected the quiver in the voice, or parsed the strained phraseology, and automatically alerted the authorities. The blue ink scratched its way across the paper, as if hard pressed to recall the individual shapes of letters. At one point the nib had pierced the white sheet. Few people wrote regularly with pens. It was still taught at school, but the odd love letter or shopping list was as far as most people got. And suicide notes, of course. This was no different; the writing was that of the very old, or the very young.

In a way the hand was old, the oldest that had still lived. But just as the sunrise is as old as time and new each dawn, so this hand was new: three months and twelve days, according to the factory's records.

Even the words, the symbols of the man's thoughts, were not worthy of note. They would have won no literary prize; inspired no doomed, romantic quest; enquickened no tired and demoralized army. The very human story was the usual one: of love, of ennui and, ultimately, of heartbreak.

No one, least of all himself, remembered quite when or how he had lost his first hand, more than 300 years ago. The accident was recorded, but if the loose-leaf binder still existed, the cheap ink was long faded into obscurity. Sometimes he claimed it was an explosion in a fume hood; at other times a gas cylinder had fallen from its moorings and crushed him.

What his memory was clear on, and what was attested to in the medical literature, was that he had attached ('single-handedly, haha!' he would joke) an artificial limb to the remains of his own arm. Not a simple *prosthetic*, but a fully functioning organ of composite fibre, ceramic joints and golden threads carrying two-way nervous traffic. The body's own electrical impulses provided power to the tiny servos that drove the slender titanium flexors and extensors.

No accident, the second prototype: it was tested and retested, planned months in advance. His wife directed the operation, and when he woke, his right arm to the shoulder was fully robotic. A fortnight later, while he was still delirious from antibiotics and analgesic, she was killed by a drunk-driver.

The record shows that he opened a new lab with venture capital, employed three dozen scientists and disappeared into his research. The exclusive clinic followed: he himself was its first patient, walking out on legs of alloyed titanium — and straight back into the lab.

Half a dozen more clinics started up across the nation, opening their doors to anyone whose medical insurance would pay the fees. For ten years the company replaced natural limbs with artificial constructs that were functionally equivalent to the original. More than equivalent: these never wore out, never got cancer, never got tired, never felt weak or cold.

For ten years the clinics operated and the lab researched. No papers were published, no patents applied for, and investors grew nervous. Interest waned. Two clinics closed; a third of the research staff was laid off. Rumours circulated, created by and lost in the noise of the Internet. It was another three years later when, finally, a press conference was called on the lawn of the first clinic, the handful of journalists who bothered to turn up were turned away — — and were called back, to face a man who under crepusculine clouds *glistened*.

The patents and the papers followed on the morrow: the artificial blood, the fuel cells, the intricate and minuscule fibres and vessels and motors: in short, a body wonderfully and fearfully man-made.

Only his face appeared natural, and over the following years even that was slowly replaced. Having no need of food, depending solely on a defined and especially formulated medium, protected by filters and powered by the elements, no toxins could threaten him. With hard, durable alloys and man-made composites in place of bones and tissues, redundant systems and every organ replaceable, he was all but indestructible.

Alzheimer's had been cured by the time he reached 105, and the last bastion of mortality — the uncontrolled cell division leading to legion neoplasms — tamed a few years after that. And then he was a living brain in a metal and plastic shell, talking, walking and living: never fatigued, immune to all disease, the Tree of Life incarnate.

For 200 years he lived like this, never needing to eat: a weekly cocktail of nutrients and pharmaceuticals keeping the one, irreplaceable fleshly and uniquely human organ alive.

When the end came it was without fanfare or press conference. No papers were written, no patent lawyers notified. With the finest of Torx drivers he opened an access panel, removed a wire, took out a power cell, held it — his life in his own hands.

The suicide note of the world's first immortal ended simply enough:

I cannot live without her. ■

rpg is the nom de plume of a molecular cell biologist and hopeless romantic at the University of Sydney.